HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Janne Argillander

Maximum Entropy Modeling and Semantic Concept Detection

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, April 29, 2005

Supervisor:	Professor Unto K. Laine
Instructor:	Dr. Giridharan Iyengar

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Janne Argillander
Name of the thesis:	Maximum Entropy Modeling and Semantic Concept Detection
Date:	April 29, 2005 **Number of pages:** 62
Department:	Electrical and Communications Engineering
Professorship:	S-89
Supervisor:	Professor Unto K. Laine
Instructor:	Dr. Giridharan Iyengar

The growing amounts of multimedia data available to the average user has reached a critical phase where methods for indexing, searching, and efficient retrieval are expressly needed to manage the information load. For instance, the amount of multimedia content that is already present in the consumer hard-drives makes manual annotation (and consequently, indexing using high-level keywords) impossible. In this thesis, we approach this problem by proposing a Maximum Entropy based approach for automatic annotation of multimedia content.

In this thesis, we present a generic and scalable approach for modeling semantic concepts in multimedia data incorporating cues from multiple modalities in a unified framework. In particular, we model the joint multimodal feature observations using the Maximum Entropy (MaxEnt) approach. We propose a set of novel predicates for both the visual features and speech features. Experiments indicate that approach is promising and its performance is comparable to that of the state-of-the-art systems benchmarked on the TRECVID corpus. In addition, the multimodal models significantly outperform (by 10 percent) unimodal approaches on this corpus.

Keywords: Maximum Entropy, MaxEnt, Concept Detection, Automatic Annotation

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Janne Argillander
Työn nimi:	Maximum Entropy Modeling and Semantic Concept Detection
Päivämäärä:	29.4.2005 **Sivuja:** 62
Osasto:	Sähkö- ja tietoliikennetekniikka
Professuuri:	S-89
Työn valvoja:	professori Unto K. Laine
Työn ohjaaja:	Dr. Giridharan Iyengar

Saatavilla olevan multimedian määrä on kasvanut niin suureksi, että sen hallinta ilman apuvälineitä on mahdotonta. Tätä hallintaa helpottavat hakumenetelmät, joiden avulla voidaan siirtyä multimediaesityksessä haluttuun kohtaan. Nämä hakumenetelmät perustuvat olemassa oleviin indekseihin, jotka on luotu käsin. Nykyään keskimääräisen käyttäjän kotikoneelta löytyy kuitenkin niin paljon multimediatiedostoja, että niiden manuaalinen läpikäyminen olisi liian työlästä. Tässä työssä haen ratkaisua tähän ongelmaan esittelemällä menetelmän, joka kykenee indeksoimaan multimediatiedostoja automaattisesti.

Työssäni esittelemäni menetelmä käyttää hyödykseen sekä visuaalisia, että puheeseen perustuvia vihjeitä. Nämä vihjeet esitetään tilastolliselle maksimi-entropiaprosessille predikaattien avulla. Menetelmän suorituskyky on suoraan verrannollinen näiden predikaattien toimivuuteen. Tämän vuoksi predikaattien suunnittelu on yksi tämän työn keskeisimmistä kohdista.

Tehdyt kokeet osoittavat, että multimodaalinen menetelmä toimii paremmin, kuin yhtä modaliteettia käyttävät menetelmät. Vertailu paljastaa myös, että esitetty menetelmä toimii vastaavalla tasolla TRECVID kilpailun voittaneen menetelmän kanssa. On myös huomioitava, että esitetty menetelmä on geneerinen ja TRECVID kilpailussa olleita menetelmiä huomattavasti yksinkertaisempi. Tämän vuoksi esitetty multimodaalinen menetelmä on lupaava ja jatkotutkimuksen arvoinen.

Avainsanat: Maksimi-entropia, MaxEnt, Semanttinen konsepti, Automaattinen annotaatio

Acknowledgements

This Master's thesis has been carried out at the department of Human Language Technologies at the IBM Thomas J. Watson Research Center.

First of all, I want to thank my thesis instructor Dr. Giridharan (Giri) Iyengar for his endless support and guidance in all fields. I learned a great deal about research in general, experiment setup, handling large multimedia data sets and controlling experimental bias. Secondly, I wish to thank my supervisor Professor Unto K. Laine for his valuable comments and feedback. I would also like to thank IBM Thomas J. Watson Research Center, David Nahamoo, Ganesh Ramaswamy and IBM Finland for their understanding, financial support and possibility to write this thesis. Finally, I would like to thank my fiancée Minka for everything, including encouragement and support in all possible ways.

Otaniemi, April 29, 2005

Janne Argillander

Contents

Abbreviations

AP	Average Precision. One method for measuring the efectiveness of a classifier.
ASR	Automatic Speech Recognition.
BOU	Best Of Unigram.
FL	Feature Loader. Type of experiment that san hold one or many types of predicates.
GIS	Generalized Iterative Scaling. A Method used for solving the Lagrance multipliers.
IBM	International Business Machines.
ICASSP	IEEE International Conference on Acoustics, Speech, and Signal Processing.
IEEE	Institute of Electrical and Electronics Engineers.
Lab	Lab (or L*a*b) is theoretical color space. This space is not tied to any output device.
MAP	Mean Average Precision. The mean AP over many concepts.
MaxEnt	Maximum Entropy.
MM	Multimodal.
MP	Mean Precision. This is mean over many concepts.
MPEG	Moving Picture Experts Group.
NIST	National Institute of Standards and Technology.
NLP	Natural Language Processing
POS	Part Of the Speech.
QBE	Query By Example.
QBK	Query By Keyword.
TREC	Text REtrieval Conference.
TRECVID	TREC Video Retrieval Evaluation.
VHS	Video Home System.

List of Figures

List of Tables

Chapter 1

Introduction

The amount of multimedia content available for consumption, especially video, is increasing at a phenomenal rate with the advent of ubiquitous recording devices such as camera phones and camcorders. We have already reached a critical point where methods for indexing, searching, and efficient retrieval are expressly needed to manage the information load. The amount of multimedia content that is already present in the consumer hard-drives makes manual annotation (and consequently, indexing using high-level keywords) impossible, mainly because both video and audio browsing is hard. There has been some effort in using query-by-example (QBE) to seek into multimedia content (e.g [36, 8] amongst many others). While QBE is a powerful paradigm, its reliance on low-level perceptual properties is counter to the *semantic* nature of most user queries[36, 38]. In this thesis Query-by-keyword (QBK) [27] system is built. In the QBK paradigm the user queries the content using semantic descriptors. Because this behaviour is more natural to human are these systems getting more and more attention. QBK systems typically require at least two processing stages: A primary *training* phase where the system is taught to identify specific concepts from a pre-defined vocabulary; A secondary *annotation* phase where the system (semi-)automatically annotates previously unseen content with these newly learned concepts. These annotations can be used in the final multimedia *retrieval* system. In this thesis the interest is in a general framework using cues from multiple modalities in order to build classifiers for automatic annotation and retrieval of semantic concepts from the multimedia data.

1.1 Annotation

In the annotation task, either human manually or computer automatically identifies useful objects, sites and events from specific content and then describes the content using key-

words to represent identified meta-data. Manual approach has many disadvantages: it is expensive, very time-consuming and generated keywords may be subjective, to name a few. Therefore, there is need for automatic systems that can attach content- and concept-based keywords to multimedia data. Note the distinction between annotation and object detection. In annotation, it is sufficient if the image containing the object is marked whereas in detection, there is a further assumption that the location, scale, and when possible, cardinality of the object of interest are correctly identified.

The granularity of annotations depends highly on the type of the index to be created and the target usage of the index. While a personal video catalogue (Table 1.1) or an library index (Table 1.2) may be fairly simple and small (granularity is small), a digital index (Table 1.3) can be relatively big and complicated. The tables illustrate the dependence of the accuracy of retrieval task on the quality of the index. It is hard to say whether any of the videos shown in personal video cataloque will contain pictures about pyramids. The situation is slightly better in the case of a library index because it tells that specific tape will contain something about pyramids. However, because it does not contain the actual position of the pyramid in the tape, the only way to check where it occurs and what is being said about the pyramids is to watch entire document.

Cassette	Description
1001	Saturday morning animations 2003-2004
1203	Movie: Lorenzo in Africa
2702	Series: Old and lazy, episodes 540, 541, 542
...	...

Table 1.1: Personal video catalogue

Code	Explanation	Value
0001	Title	Pyramids. Who made them?
0002	Code	VID199.4483200
0300	Format	VHS Cassette, 145 minutes
1000	Keywords	Sphinx, Pyramid, Egypt, Nile
...	...	...

Table 1.2: Library video index

Digital index offers enough information to seek immediately into specific locations of the video content. In order to build a digital index automatically, video content (which in this thesis is always in MPEG format [15]), has to be presented in smaller units than documents or films. One of the most popular approaches for performing this is a structured modeling approach [6]. In this approach video content is first divided into *shots* using automatic shot

Tag	Value
Video	20042507_ABC
Start frame	0
End frame	260
Key-frame	134
Start time	00:00:00:00
End time	00:00:08:21
Key-frame time	00:00:04:15
Annotation	Pyramid, Sand, Human, Face, Outdoors, ...
...	...

Table 1.3: Digital video index

boundary detection system (cf. [14]). A shot is an image sequence coming from single operation of camera, and presenting continuous action. This image sequence can be further divided into single images called frames. In this thesis the level of annotation is on the shot level. This means that the task is to find out the probability for specific concept given current shot $p(concept, shot)$.

Given a shot, either multiple frames in the shot can be processed or, to reduce computational complexity, a representative keyframe can be processed by the automatic annotation system. In this thesis, the focus is primarily on keyframe based systems. After keyframe extraction multiple features such as co-occurence, LAB-moments, smoothed edge angle histograms and automated speech recognition (ASR) transcripts are extracted for each video shot. Based on these features, automatic annotation process should be able to determine keywords for every shot. An example image (Fig. 1.1) and some possible annotations are shown next:

- Scenes (Background): Outdoors, Building, Sky,...
- Objects/Visual: Tree, Person,pier,...
- Objects/Auditory: Music, Speech, Sound,...
- Activities: Leisure, Sun worshipping, walking,...

Unfortunately the automatic annotation is not a trivial task. Size, color, position, angle, used words and other circumstances of the interesting object or background may vary a lot between the scenes. Different objects have different properties like one is rounded and one is angular, and the ASR words spoken while object present may be arbitrary. And, some objects or entities (like in the natural world) have no characteristic features that are sufficient for discriminating specific concept [37]. We have to deal with many cues and

Figure 1.1: *One frame from the shot taken in Central Park*

modalities which will lead easily to very high dimensional feature vectors. The problem is to know which feature or set of features are best to describe specific object or concept.

The annotation task can be seen as a signals-to-symbols paradigm [37] where one tries to fit something like all possible trees in the world into common "tree class". And the process where we try teach a machine to do this fitting for us is called machine learning [35]. All this leads into the fundamental question, "what makes a tree to be tree". And is it possible to detect a tree without detecting the forest first, but do we have to detect tree first before forest? Or, should we try to detect everything at the same time. Most likely it would be ideal to learn something from the human perception. Unfortunately, the study of human perception is not in such level that it would give straight answers [11].

1.2 Related work

There is extensive literature in object detection (especially human faces) where the spatial extent of an object is well-marked in an image. There is relatively limited literature in automatic image annotation where the physical extent of the objects are not specified. In one set of approaches, techniques from statistical machine translation were applied to the problem of image annotation. In these approaches it is assumed that the annotation and the associated image are translations of each other and with a suitable of *tokenization*

of the image features, standard machine translation models have been applied with some success[31]. Motivated from a cross-lingual information retrieval perspective, Lavrenko et al.[22] approach image annotation as an example-based learning problem where perceptual similarity in the image space is assumed to generate similar annotation words. Both these approaches have been demonstrated on relatively small datasets (5000 images from COREL dataset) and they remain to be evaluated in larger contexts such as what is attempted in this thesis (e.g. 80000 shots from TRECVID2003 corpus[28]). These visual results are also published (please see [17] or Appendix A, Chapter 8).

Motivated by the under-constrained nature of the annotation problem together with the non-independent nature of low-level image features, the Maximum Entropy (MaxEnt) setting which has had remarkable success in many Natural Language Processing tasks such as sentence-boundary detection and parts-of-speech tagging[34, 39] is selected for the modeling task. A similar approach using MaxEnt for image annotation was proposed in[20]. The novelties of the approach used in this thesis are two-fold: The spatial- and joint-dependence between low-level features using specially designed predicates is modeled. We strongly believe that such information is important for objects that have a well-defined spatial composition (e.g. faces). In addition, the evaluation is made on a much larger corpora (TRECVID2003 and 2004). The comparison between approach presented here and with previously published results on the TRECVID2003 concept detection task[28, 3] is also made.

1.3 Thesis objectives

In this thesis the area of interest is in a general framework using cues from multiple modalities in order to build classifiers for annotation and retrieval of semantic concepts from the multimedia data. The classifiers are first trained by using a training corpus of human annotated examples and then used to annotate new unseen multimedia such as images and videos. For the training, one speech and three types of low-level image features are extracted. These features are tokenized and fed into Maximum Entropy model. The detailed process is defined in following chapters. The performance of these classifiers is evaluated using a test set. The corpora that we choose to train and test the classifiers is the newly established TRECVID2003 and TRECVID2004 datasets. These are the largest known publicly available multimedia corpora for which well established benchmarking and ground truth is available. The final aim is to build fairly simple in complexity but as good in performance as the best state-of-the-art performing classifier.

Chapter 2

Dataset and Measures

2.1 Dataset

In this thesis the TRECVID2003 and 2004 [28, 30] corpora comprising 120+65 hours of broadcast news videos are used for experiments. Corpora are provided by National Institute of Standards and Technology (NIST). These are divided into approximately evenly sized test and development partitions. For the development partition NIST has provided ground truth annotations at the video-shot level. It has to be noted that these annotations are provided at the shot-level and do not specify spatial or temporal boundaries of objects within a shot (i.e. we know that a face appeared in the shot but we do not know *when* and *where* within this shot). In order to train and validate our classifiers, we further divide the development set into the following sets: training, validation, concept fusion 1 and concept fusion 2. This is done so that there would be separate sets for the system validation and for the model fusion. Finally, the trained and validated classifiers are evaluated on the test partition and compared against NIST provided relevance judgements. The comparison with relevant competing systems, when appropriate, is also provided. The videos are stored as MPEG-1 files and the meta-data is stored in MPEG-7 format. In the year 2004, the TRECVID benchmark workshop had 33 finishing participants from various universities, and research organizations.

TRECVID benchmark has four main tasks:

- Shot boundary detection
- High-level feature extraction
- Story segmentation
- Search (IR) task

In this thesis the general focus is on the High-level feature extraction task which is similar to automatic annotation as we defined earlier. The main difference is that in automatic annotation the decision to annotate or not is based on some threshold possibly determined by acceptable precision and recall criteria for the task at hand. If the predicted probability is above this threshold then the shot will be annotated with this keyword, otherwise not. In the TRECVID High-level feature extraction task, all shots are annotated with the probability of the concept. The probabilities are ordered and then ranked. The goodness of a particular classifier comes from the ability to return positive shots at the top of the list. Because of sorting there is no need to define any thresholds.

In order to facilitate standardized evaluation, NIST extracts two fundamental information sources from the data: *keyframes* and *ASR words* which are used by all groups participating in tasks other than Shot boundary detection. In this thesis a total of 32 binary classifiers, one for each concept, is built. The list of concepts for which the classifiers are built and evaluated (see Table 2.1) are the concepts that were benchmarked in the TRECVID2003 and 2004 workshops.

2.1.1 Human labeled annotations

It is good to have some training data with high-quality, manually prepared annotations. Unfortunately, due to unsurmountable subjective biases amongst human labellers, such annotations are rarely noise-free, except for a very small well-defined vocabulary of concepts. Some such manual labelling problems are briefly illustrated below.

Many of the shots are *missing* annotations. That is, while the concept is present in the shot, the labeller has not marked the corresponding keyword in the annotation. This does not affect training but when evaluating on the development set, these are treated as incorrect retrievals.

Spelling of some annotations is incorrect. The same keyword has been written in multiple ways. Example: "Courtoom" and "Courtroom" , "Missle" and "Missile" , etc. It is important to note that this can be mitigated with a good choice of annotation interface where the human labeller selects the word from a pre-determined vocabulary as opposed to typing a keyword on their own.

The specific concept is not visible/audible at the point where the key-frame is extracted. Annotations are prepared on the shot level. There is only one extracted key-frame per shot which means that a specific concept may or may not be active exactly at this point. This adds some extra noise to the model because in this thesis the classifier is visually trained only based on the key-frames. However, it seems that this is not as bad as it sounds. This is because many of the annotators were "lazy" and based their decision to annotate a specific keyword by looking at just the key-frame and not the entire shot.

concept	positive
Airplane	96
Animal	157
Basketball	92
Beach	23
Bill Clinton	64
Boat	37
Building	407
Car	348
Face	4015
Female Face	1188
Female Speech	2272
Hockey	38
Indoors	3080
Madeleine Albright	7
Male Face	1998
Male Speech	4503
Man Made Scene	844
Nature Non-Vegetation	1029
Nature Vegetation	521
News Subject Face	417
News Subject Monologue	312
Non-Studio Setting	2129
Outdoors	2474
People	1214
Person Action	2762
Physical Violence	44
Road	223
Sky	597
Sport Event	308
Studio Setting	656
Train	24
Weather News	41

Table 2.1: Concepts and corresponding number of positive example shots in the training set.

Some of the shots are *incorrectly* annotated. This is the worst case because these shots will distract the modeling process.

The exact spatial or temporal *location* is not in the annotations. The annotations used in this thesis only contain the active concepts. If something is not active it does not say if it is

not really there. And, even if something is active the actual location is not known. Modeling process should be able to learn how the face looks like even if it has no knowledge what was the face in the positive shot. This is identical in spirit to the Multiple Instance Learning problem [25].

Some of the annotations seem irrelevant. For example, one annotation from one shot contains "Anchor intro and reporter voice over for Ken Starr in a car".

It has to be noted that it is a different task to build models based on annotations prepared by the experimenter than prepared by the group of different people. When experimenter defines the keywords he/she has a common criteria to choose whether to annotate a particular shot with a specific keyword or not. This introduces a subtle bias into the learning process and makes data set very stable for his/her use. Learning task is much harder when the annotations are prepared by a larger group of people. In this thesis, the annotations were provided by NIST and it was based on a collective effort of over 100 individual annotators. For example, if we show the Figure 1.1 to group of annotators and ask them to add keywords. Is is obvious that all annotators won't come up with the same annotations.

2.2 Evaluation Measures

The performance of a particular model is measured against itt's effectiveness to retrieve positive examples at the beginning of ranked list. This leads to following definitions of evaluation criteria[29]:

The *precision* P_r is a value that measures the percentage of retrieved documents r so far that are relevant to the concept.

$$P_r = \frac{\text{number of relevant retrieved}}{\text{total retrieved}} \tag{2.1}$$

The *recall* R_r measures the percentage of relevant documents in the corpus that have been retrieved so far.

$$R_r = \frac{\text{number relevant retrieved}}{\text{total number of relevant documents in the corpus}} \tag{2.2}$$

Both precision and recall are set-based measures because they evaluate the quality of an unordered set of retrieved documents. The *precision-recall graph* $P_r(R_r)$, shows the interaction between recall and precision at the point r (i.e. r documents retrieved) and therefore evaluates ranked lists.

The *interpolated precision recall curve* is developed to enable averaging and performance comparison between different systems. The precision at the recall point λ is the maximum

precision at or after the recall point λ. See NIST definitions [29] for the details on these measures.

$$P_\lambda = max(P_r) \text{ where } r \geq \lambda \text{ and } \lambda = 0.0, 0.1, ..., 1.0 \tag{2.3}$$

The averaging over several systems is computed by summing P_λ of all systems at the specified recall λ and then dividing by number of systems.

$$P_{\lambda a} = \frac{\sum_{systems} P_\lambda}{\|\text{systems}\|} \text{ where } \lambda = 0.0, 0.1, ..., 1.0 \tag{2.4}$$

The *average-precision* AP is the average of all precisions over all relevant documents. The relevance of the first retrieved documents is more significant to the AP value than latter ones. If AP has to be calculated after n retrieved relevant documents then it is possible divide the precision sum over relevant documents with the minimum value of relevant or retrieved documents. In the TRECVID2003 AP was determined after 1000 documents retrieved. And in the year 2004 after 2000 documents retrieved.

$$\text{AP} = \frac{1}{\min(\text{relevant}, \text{retrieved})} \sum P_{relevant} \tag{2.5}$$

2.2.1 Measures in Action

For example, let us assume that there are a total of 20 documents in the database. From the 20 documents, 8 are relevant. If the first retrieved document is relevant then precision at 1 is $P_1 = 1$, recall is $R_1 = 1/8$ and average precision $AP = 1$. The ideal classifier gives all the relevant document first. Therefore, for the ideal classifier, the precision is one until recall is one. The worst possible classifier gives all non-relevant documents first and then the relevant ones. This time precision and recall is zero and slowly increases in the end of list. Most precision-recall curves for classifiers lie between these two extremes. Table 2.2 shows an example. This example is also visualized in Figure2.1.

r	relevant	P_r	R_r	AP
1	+	1	$1/8$	1.000
2	-	$1/2$	$1/8$	0.500
3	+	$2/3$	$2/8$	0.556
4	+	$3/4$	$3/8$	0.604
5	+	$4/5$	$4/8$	0.643
6	-	$4/6$	$4/8$	0.536
7	-	$4/7$	$4/8$	0.460
8	+	$5/8$	$5/8$	0.480
9	+	$6/9$	$6/8$	0.564
10	-	$6/10$	$6/8$	0.564
11	+	$7/11$	$7/8$	0.643
12	-	$7/12$	$7/8$	0.643
13	-	$7/13$	$7/8$	0.643
14	-	$7/14$	$7/8$	0.643
15	-	$7/15$	$7/8$	0.643
16	-	$7/16$	$7/8$	0.643
17	+	$8/17$	$8/8$	0.702
18	-	$8/18$	$8/8$	0.702
19	-	$8/19$	$8/8$	0.702
20	-	$8/20$	$8/8$	0.702

Table 2.2: Precision-Recall example where r stands for number of documents retrieved, P_r is precision, R_r is recall and AP is average precision.

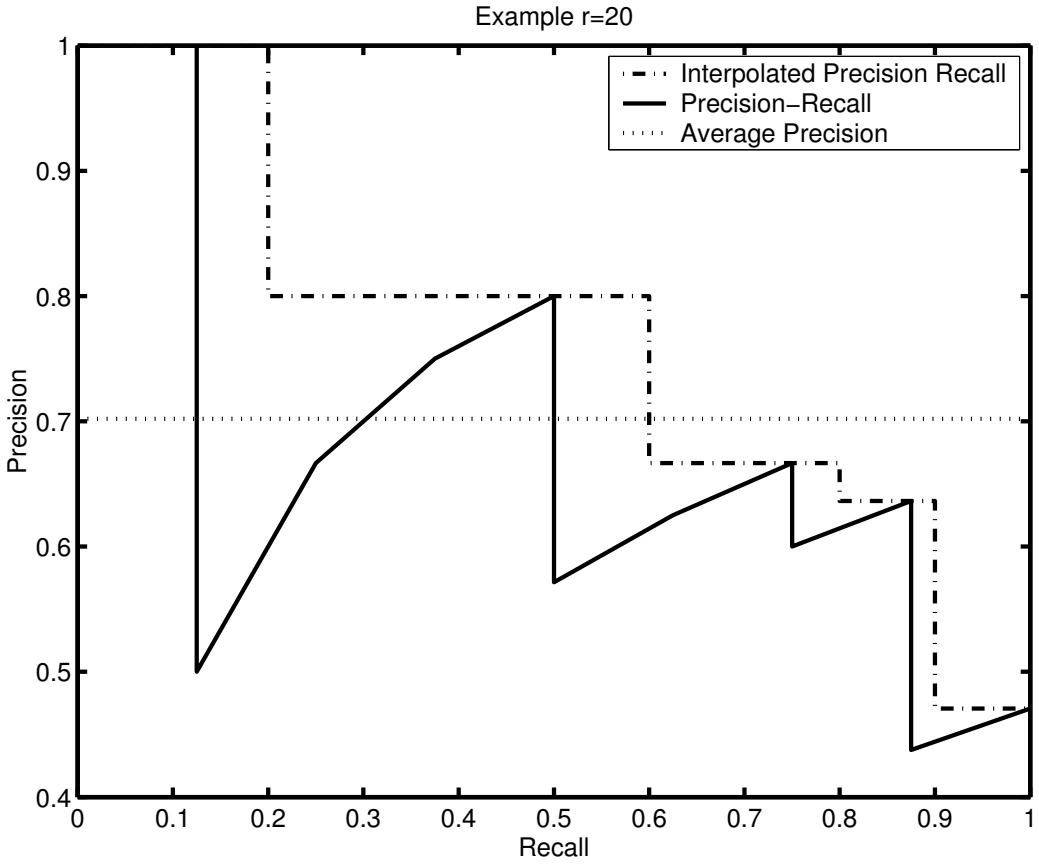

Figure 2.1: *Figure illustrates the difference between three measures.*

Chapter 3

Maximum Entropy Approach

3.1 Introduction

Let us assume that there is a random process that produces an output y given a an observation (also known as *context*) x. In multimedia annotation, y, which is a member of a finite set (vocabulary) Y, can be seen as a label for a specific shot/image. And x, a member of a finite set X, as extracted information (features) from the current shot. In this thesis training data for modeling process is presented in pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. The task is to learn possible correlations between x and y, and to build statistical models that can be used to annotate previously unseen shots automatically. This can be seen as supervised data-driven learning. The empirical probability distribution function (pdf) based on training data is as follows

$$\tilde{p}(x, y) = \frac{1}{n}\text{freq}(x, y) \tag{3.1}$$

Where *freq* is the count of a specific pair (x, y) and n is the size of the training data. In real world applications, training set size is finite. Therefore, the empirical distribution is a poor estimate of the joint pdf. The question arises, how to handle missing parts of the training in order to estimate pdf that generated empirical distribution $\tilde{p}(x, y)$ in reasonable way?
For example, let us assume that there is a players dice which has 10 sides. The task is to find probabilities for all sides (a discrete pdf). It is obvious that there is an infinite number of ways to assign these probabilities. However, due to the lack of observed information, the most uncertain, unbiased way to assign these probabilities is to choose the uniform distribution as the target pdf.

In the real world applications, pdf is not likely to be uniform. Therefore, based on observation it is possible to add some constraints into the mix. In this case it means that you

x	1	2	3	4	5	6	7	8	9	10	$\sum p(x)$
p(x)	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	1

Table 3.1: The most uncertain solution

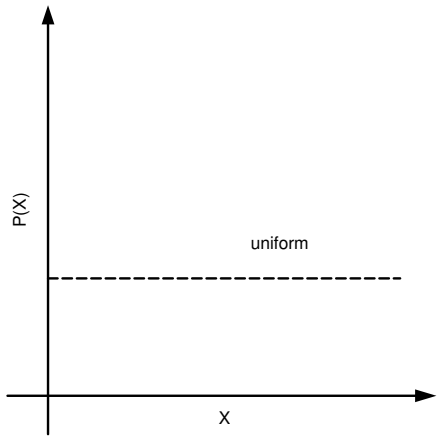

Figure 3.1: *The uniform solution is the most uncertain solution*

observe which players are not fair. Someone may have weighted his/her dice so that the side 3 has more weight than others for example. However, if we have observed this "feature" it is possible to refresh our probability for the side 3. And, again because we don't have any other information such as knowledge from the other sides we better use uniform solution for other probabilities.

x	1	2	3	4	5	6	7	8	9	10	$\sum p(x)$
1	$1/20$	$1/20$	$11/20$	$1/20$	$1/20$	$1/20$	$1/20$	$1/20$	$1/20$	$1/20$	1

Table 3.2: The most uncertain solution in the case of constraint

3.2 The Maximum Entropy principle

Let us now formally state the maximum entropy principle.

> Two events should be assigned with equal probabilities if there is no reason to think otherwise.
>
> Laplace - "Principle of Insufficient Reason"

A related example of such thinking is the Occam's razor or least complex hypothesis selection. Laplace can be considered as a father of Maximum Entropy (*MaxEnt*) [19, 18].

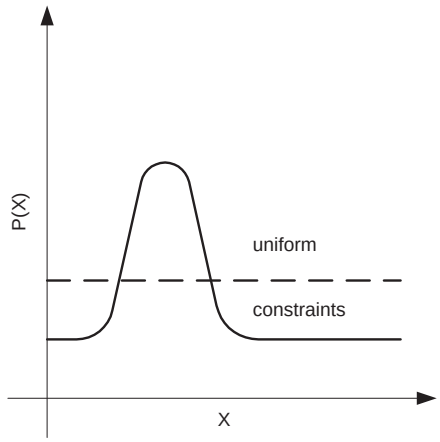

Figure 3.2: *It is good to follow uniform solution when no better knowledge is available*

His "Principle of Insufficient Reason" was an first attempt to apply an unbiased method into field of probability theory. However, the practical formalism and the answer to the fundamental problem was first introduced by the E.T. Jaynes [18] and then updated by I.J. Good [13]. Jaynes suggested that the most unbiased selection of the pdf can be made by maximizing Shannon's entropy (Formula 3.7).

$$H(X) = -\sum p(x) \log p(x) \tag{3.2}$$

This leads to the distribution that has maximum entropy, maximum uncertainty, which is the most unbiased and most uniform. At the same time the MaxEnt principle keeps the pdf consistent with the observed constraints. These constraints are variously called as *feature functions*, *feature indicators* or *predicates*. Specifically, feature function is a function that gets a value greater that zero if the (contextual) predicate(s) defined by the feature function can be observed. In this thesis, feature functions in the binary form are used. There is no difference in value between active (contextual) predicate or feature function. Therefore, In this thesis the term *predicate* over *feature function* is preferred to avoid confusion with the extracted low-level features.

3.3 Predicates

At the core of the modeling process are the predicates. These are used to specify constraints on the model. In MaxEnt, the process of defining predicates is central to modeling: The goodness of the resulting models is dependent on the ability of these predicates to capture the relevant information. All predicates used in this thesis are binary-valued. This means

that predicate either fulfills its definition or not. By the means of predicates it is possible to force the model to agree with some statistics (as in the previous example) determined as important by experimenter. Such statistics in multimedia annotation is usually the extracted knowledge from the context. For example,

$$f(context, label) = \begin{cases} 1 & \text{if } label = \text{sky and } cd = \text{blue} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

This predicate is active only if it color blue is observed and this specific shot is annotated as a sky by human.

3.4 Model Preparation

In this thesis, a distinct binary classifier for each semantic concept is built. In this thesis, the multimedia annotation problem is set up similar to the Multiple Instance Learning problem. In the Multiple Instance learning approach [25], each labeled example containing the target concept annotation is considered as a positive example for that classifier. All training instances that do not contain the target annotation are marked as a negative examples. It has to be noted that each training example has multiple concept labels in its ground truth annotation. E.g. A shot may be annotated as a face, outdoors, sky etc. This imposes the possibility to built classifiers that conflict with each other. E.g. every shot annotated as a "moon" may not be annotated as a sky. This means that even if the "moon" classifier says that there is moon in the shot, the "sky" classifier may not agree with that. Given training data, we can calculate the expected value of a particular predicate:

$$\tilde{p}(f) \equiv \sum_{x,y} \tilde{p}(x,y) f(x,y) \tag{3.4}$$

These empirical expectations of the various predicates provide constraints for the target concept model.

$$p(f) \equiv \sum_{x,y} \tilde{p}(x) p(y|x) f(x,y) \tag{3.5}$$

$$p(f) = \tilde{p}(f) \tag{3.6}$$

The maximum entropy principle states that the unbiased choice is to select the most uncertain (uniform) distribution from a family of distributions $p(y|x)$. This uncertainty can be measured in an information theoretic sense using Shannon entropy as follows:

$$H(p) = -\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \tag{3.7}$$

The minimum uncertainty is zero which yields a perfect predictor. The upper bound for the uncertainty is $log(|Y|)$ when the chosen distribution is completely uncertain (uniform). The model with maximum entropy (given the empirical constraints) is:

$$p_* = \arg\max_p H(p) \tag{3.8}$$

Barring trivial cases, it is impossible to directly solve for this equation. The classical approach is to introduce a Lagrange multiplier λ [5]:

$$\Lambda(p, \lambda) \equiv H(p) + \sum_i \lambda_i (p(f_i) - \tilde{p}(f_i)) \tag{3.9}$$

With λ fixed it is possible to compute the maximum of the Lagrangian $\Lambda(p, \lambda)$. The point $\Psi(\lambda)$ where $\Lambda(p, \lambda)$ achieves its maximum is denoted by p_λ:

$$p_\lambda \equiv \arg\max_\lambda \Lambda(p, \lambda) \tag{3.10}$$

$$\Psi(\lambda) \equiv \Lambda(p_\lambda, \lambda) \tag{3.11}$$

Expanding Eq. 3.9, we get

$$\begin{aligned} \Lambda(p, \lambda) = &- \sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \\ &+ \sum_i \lambda_i \Big(\sum_{x,y} [\tilde{p}(x) p(y|x) f_i(x, y) - \tilde{p}(x, y) f_i(x, y)] \Big) \\ &+ \gamma \sum_x p(y|x) - 1 \end{aligned} \tag{3.12}$$

Where the first term is the entropy (Eq. 3.7) and the second term comes from substituting the definition of $p(f)$ from Eq. 3.5. We introduce the last term and the constraint (γ) to ensure that $p(y|x)$ is a valid probability density function. Solving for the maxima of this quantity with respect to $p(y|x)$, we get (this is also too big leap for me)

$$[- \sum_{x,y} \tilde{p}(x) \log p(y|x) - \sum_{x,y} \tilde{p}(x) + \sum_i \lambda_i (\sum_{x,y} \tilde{p}(x) f_i(x, y)) + \gamma \sum_x 1] \times \frac{\partial p(y|x)}{\partial x} \tag{3.13}$$

Considering only terms that matter (the external partial derivative does not matter) and noting that $\sum_{x,y} \tilde{p}(x) = 1$ and $\gamma \sum_x 1$ is some constant, we get the following simplified form.

$$- \sum_{x,y} \tilde{p}(x) \log p(y|x) + \sum_i \lambda_i (\sum_{x,y} \tilde{p}(x) f_i(x, y)) + C \tag{3.14}$$

where we introduce the constant C as a convenience for the various constant terms. Rearranging terms and taking exponents, we get

$$p\left(y|x\right) = \frac{\exp\left(\sum_i \lambda_i f_i\left(x,y\right)\right)}{Z_\lambda(x)} \tag{3.15}$$

$$\Psi(\lambda) = -\sum_x \tilde{p}(x)\log Z_\lambda(x) + \sum_i \lambda_i \tilde{p}(f_i) \tag{3.16}$$

The partition function Z_λ is chosen to satisfy the requiremen that p(y|x) is a valid pdf. We note that it is independent of y and depends only on λ and x.

$$\sum_y p_\lambda(y|x) = 1 \tag{3.17}$$

$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i\left(x,y\right)\right) \tag{3.18}$$

It has been shown in the literature that function $\Psi(\lambda)$ (3.16) has a straight relationship to the Maxmimum Likelihood. Thus, The model with maximum entropy is the model that maximizes the likelihood of the training sample $\tilde{p}$ [5].

$$\Psi(\lambda) = L_{\tilde{p}}(p_\lambda) \tag{3.19}$$

$$L_{\tilde{p}}(p_\lambda) \equiv \log \prod_{x,y} p(y|x)^{\tilde{p}(x,y)} = \sum_{x,y} \tilde{p}(x,y)\log p(y|x) \tag{3.20}$$

The chosen model is obtained by solving for the maximum of $\Psi(\lambda)$:

$$\lambda^* = \arg\max_\lambda \Psi(\lambda) \tag{3.21}$$

In the real world applications, an analytical solution to this function cannot be found. Instead, numerical estimation techniques are employed. In this thesis the Generalized Iterative Scaling (GIS)[7] algorithm (with smoothing) is used. GIS is not the best algorithm available for this task, but it was still selected because of its simplicity and stability.

3.5 Parameter Estimation

In order to use GIS there are some limitations that need to be adhered to. First of all, the predicates must obtain either value 1 as active or 0 as not active. Partially active predicates are prohibited. Secondly, for every event there must be at least one active predicate. Finally, the predicates must to sum to a constant C for any (x,y). This means that there must be equal amount of active predicates for every event.

$$\sum_i f_i(x,y) = C \tag{3.22}$$

In real world applications this constraint can be impossible to accommodate. Therefore, it is reasonable to add a correction predicate that maintains the constant C over arbitrary (x, y).

$$C_{corr} = C - \sum_i f_i(x, y) \tag{3.23}$$

At the beginning of the GIS iterations, all weights for all predicates are set to 1.

$$\lambda_i^{(0)} = 1 \tag{3.24}$$

Every iteration adjusts the predicate weights in the following way:

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} \left(\frac{\tilde{p}(f)}{p^{(n)}(f)} \right)^{1/C} \tag{3.25}$$

where

$$p^{(n)}(f) = \sum_{x,y} \tilde{p}(x) p^{(n)}(y|x) f(x, y) \tag{3.26}$$

$$p^{(n)}(y|x) = \frac{\exp\left(\sum_i \lambda_i^{(n)} f_i(x, y) \right)}{Z_\lambda(x)} \tag{3.27}$$

After every iteration, a log-likelihood $L_{\tilde{p}}(p^{(n)})$ of the empirical distribution $\tilde{p}$ as predicted by a model p is calculated [5]. If the change in log-likelihood, compared to previous value, is negligible then the computation will be terminated.

$$L_{\tilde{p}}(p^{(n)}) = \sum_{x,y} \tilde{p}(x, y) \log p^{(n)}(y|x) \tag{3.28}$$

In this thesis, other stopping criteria that we have used is terminating after a set number of iterations (150 iterations, determined empirically as a good value). After the weights for every feature have been estimated it is possible to do prediction/annotation using the resulting model.

3.6 Automatic Annotation of Unseen Multimedia Content

Semantic concept annotation of unseen multimedia content proceeds in the following manner. First, the low-level feature descriptors are extracted from the data. For each concept to be predicted, the set of all active predicates relevant to the concept are extracted from the feature descriptors. Now there is enough information to estimate the conditional probability of the particular concept for the shot $p(concept, predicates)$.

We now detail the visual cues and the speech cues used in this thesis in Chapters 4 and 5.

Chapter 4

Maximum Entropy Modeling Experiments With Visual Cues

4.1 Feature Preprocessing

In this thesis, three types of low-level image features are extracted from the each video shot: Lab space color moments (mean, variance, skewness and kurtosis for each channel), Edge orientation histogram (Edge strength and orientation values at each pixel, each quantized to 8 bins) and summary statistics of grey-level co-occurrence matrices (entropy, energy and contrast values). Together, these form the low-level descriptors which are termed as *Color*, *Edge* and *Texture* henceforth in this thesis. Furthermore, each shot key-frame (comprising 350×240 pixels) is partitioned into 35 regions (50×48 pixels each) and the above features for each of these 35 regions are extracted, see Figure 4.1. Features extracted from these regions are then tokenized using the K-means algorithm. In K-means, we set the distance between two feature vectors to be their Eucledian norm.

4.1.1 Unigram predicates

Unigram predicates are defined to capture the co-occurrence statistics between a specific tokenized descriptor and manual annotation of the training data. These predicates are not tied to any specific region. Therefore, these predicates have equal weight in all regions of the keyframe. All unigram predicates used in this thesis have the following form:

$$f_{cd^i,a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i \in x^i, i = 0, 1, 2 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

A predicate of this type is active only if the tokenized descriptor cd is in current frame x and the corresponding manual annotation is a. The total number of unique unigram predicates

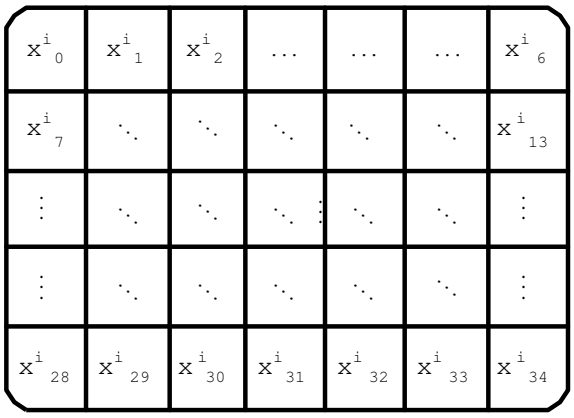

Figure 4.1: *The figure illustrates the 35-region grid partition of shot key-frames.*

used in the experiments is (descriptor count x cluster size) 3x25=75.

4.1.2 Place Dependent Unigram Predicates

Place dependent unigram predicates are designed to capture location specific statistics. For instance, these predicates help the model learn that regions corresponding to *sky* are usually in upper parts of a key-frame.

$$f_{cd^i,a}(x_r^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i = x_r^i \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

Where region r has values $0...34$ (representing the 35 regions in every key-frame) and descriptor i takes values $0, 1, 2$. The place dependent predicate is active only if the tokenized descriptor cd is in region r of the current frame which has the annotation a. The total number of such predicates used in the experiments is (descriptor count x region count x cluster size) 3x35x25=2625.

4.1.3 Bigram Predicates

In this thesis I experimented with four types of bigram predicates:horizontal, vertical, diagonal and skip-gram. These predicates model the relationship between neighboring regions. Below is an example of a horizontal bigram predicate which is active only if tokenized descriptor cd_r and its horizontal neighbor cd_{r+1} is adjacent in current frame x with annotation a.

$$f_{cd_r^i+cd_{r+1}^i,a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd_r^i + cd_{r+1}^i \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

Where the region r take values so that the adjacent region on right cd^i_{r+1} is in the row. The following equation illustrates a vertical bigram predicate which is active only if tokenized descriptor cd_r and its vertical neighbor cd_{r+7} are also adjacent in current frame x.

$$f_{cd^i_r+cd^i_{r+7},a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i_r + cd^i_{r+7} \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Where the region r take values so that the adjacent region below is in same column $r = 0...27$.

All types of bigrams are constructed by combining the tokenized features in the product space of the unigram predicates. This choice imposes the possibility of obtaining bigram values that are not supported in the training data, resulting primarily from the sparseness of the product space. To counter this, an approach inspired from class-based language models in speech processing is employed. When two unigrams are composed into a bigram, it is treated differently. We start with a small cluster of the composed unigrams. If the number of clusters is such that the number of unique bigram predicates observed (in the training data) at each step matches the total possible bigram product space values then number of cluster is increased. This process is terminated at the largest cluster size for which this condition is met in the training data.

4.1.4 Joint Observation Predicates

The predicates discussed so far model individual low-level feature descriptors (i.e. Color, Edge, Texture). Now predicates that model the interactions between the various low-level feature descriptors are illustrated.

$$f_{cd,a}(x, y) = \begin{cases} 1 & \text{if } \forall i \; y = a \text{ and } cd^i \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

This joint observation predicate is active only if all low-level descriptors are present in a given region. In the experiments, there are 144 such joint observation predicates, chosen using validation data. Figure 4.2 illustrates the various types of predicates used in the model. The diagonal and skip-gram predicates are missing from the picture. This because these predicates so far had a negative effect on the performance of the models. This can be one effect from the square-shaped regions. The distance between two adjacent diagonal regions is more than distance between two adjacent horizontal or vertical regions. It has to be noted that it is impossible to move up to n-grams (with this scheme) due to small amount of positive examples for specific concept in the training set.

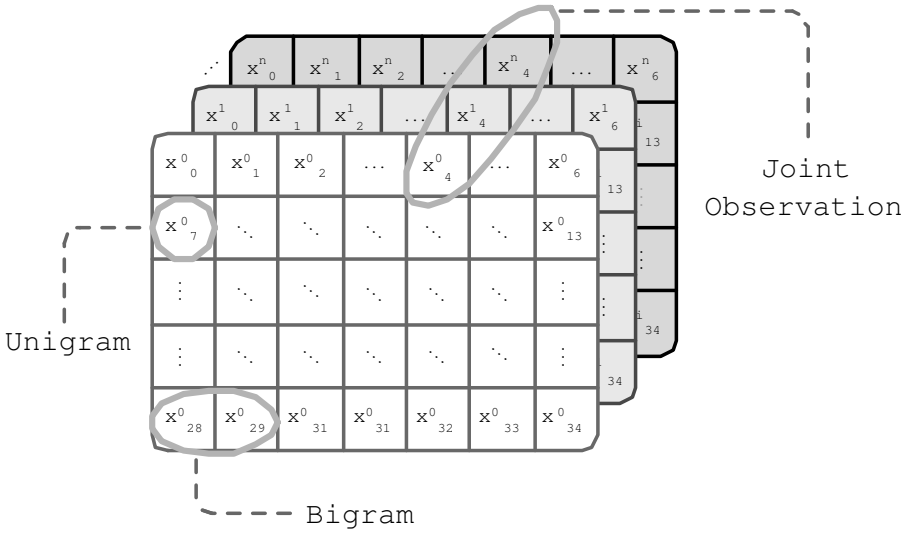

Figure 4.2: *The figure illustrates the 35-region grid partition of shot key-frames. Shown in the figure are the place-dependent unigrams, horizontal bigrams and joint-observation predicates.*

4.2 Experiments on the development set

The Table 5.1 describes three experiments conducted using the above mentioned visual predicates.

name	contents
FL-0	The visual baseline experiment using unigram predicates only
FL-4	unigram and bigram predicates
FL-21	unigram, bigram, place dependent and joint observation predicates

Table 4.1: Conducted visual experiments.

All these classifiers were built using the training set and then evaluated using the larger development test set (set concept fusion 2 in the dataset description). Specific shot is shown as a positive example of a concept if it is annotated with the concept. Otherwise, shot is treated as a negative example. As it can be seen from Figure 4.3 and from Table 4.2, the performance improves when new types of predicates are added into model. It has to be noted that larger number of predicates implies the need for big training set with many positive examples for a specific class. In another words, it is almost impossible to estimate weights for millions of predicates while having only few positive examples. In this case the maximum entropy classifier will always predict that the particular specific class does not exist (very low probability for the concept $p(concept|predicates)$). This happens because

the probability for that concept $p(concept)$ is very low (ex. $1/10000$) in the training set. In some sense, the most uncertain prediction is in this case is to always predict the null class (or background).

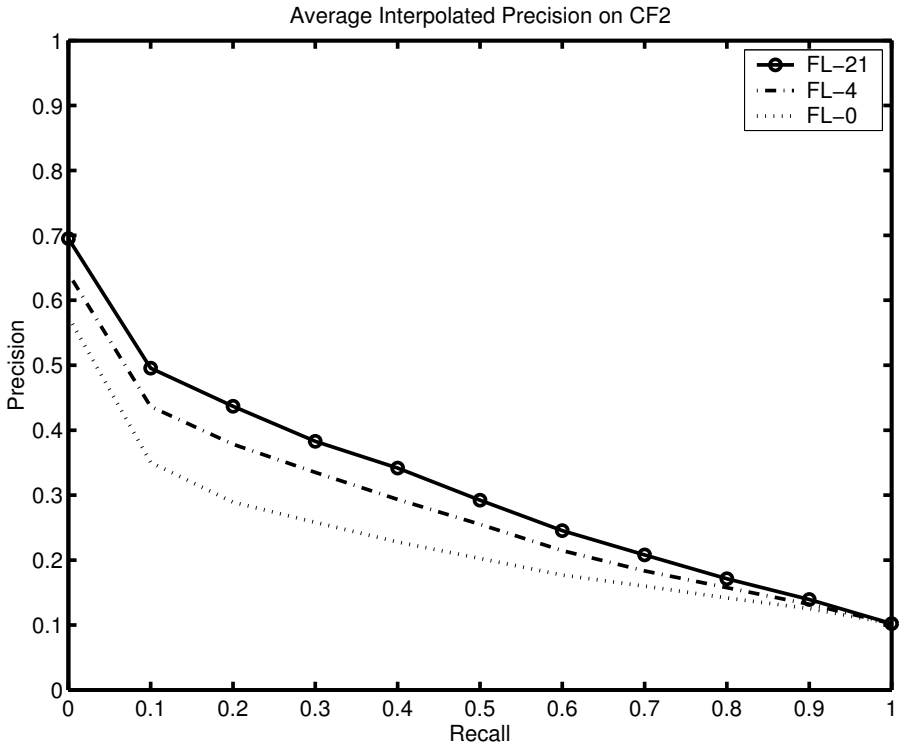

Figure 4.3: *The figure illustrates how performance is affected when more predicate types are added.*

experiment	MP@100	MAP@1000
FL-0	0.313	0.143
FL-4	0.416	0.208
FL-21	0.461	0.252

Table 4.2: The change in mean average precision and mean precision while number of different predicates varies.

4.3 Comparison against the state-of-the-art models at TRECVID2003

The performance of a system presented in this thesis is compared to the best TRECVID2003 submission with the NIST-evaluated relevance judgements reported on the test partition. As stated before, for the development partition NIST has provided ground truth annotations at

the video-shot level. In addition, NIST has provided reference key-frames for each shot for the entire corpus. For each of these reference key-frames the specified low-level feature descriptors are extracted as indicated earlier. The annotations provided are at the shot-level and do not specify spatial or temporal boundaries of objects within a shot (i.e. the knowledge that a face appeared in the shot is present but *when* and *where* within this shot is the mystery). 12 of the 17 benchmarked concepts from TRECVID2003 were selected; The audio concepts (*Female Speech*), abstract concepts (*Physical Violence*), specific person concepts (*Madeleine Albright*), camera operations (*Zoom-in*) and multimodal concepts (*News Subject Monologue*) were removed. In the case of audio and multimodal concepts, these were removed because low-level features do not capture relevant information. Camera operations do not belong in the same category of concepts as the rest of the concepts. Both *Physical Violence* and *Madeleine Albright* had very few training examples. It has to be noted that the performance of the benchmark systems on these concepts were quite low as well. For each of the selected concepts, a MaxEnt classifier using the visual cues as previously stated is built. These trained classifiers are then used to annotate the test corpus.

4.3.1 Results

NIST provides pooled relevance judgements and since this system was not part of this pooling, it would be unfairly biased to compare the system with the pooled judgements. To make the comparisons valid, evaluation between the different systems is conducted using precision at the top 100 retrieved shots as opposed to the average precision metric that is used by NIST. Furthermore, the comparison is restricted to two of the top 10 semantic concept detection systems at TRECVID2003: the best performing (multimodal) system and the best unimodal system[3]. All results are detailed in Table 4.3. The table also details the 12 concepts classifiers that were built. The first column BOU (Best Of Unimodal) is formed from the set of models by selecting the best performing unimodal classifier for the semantic concept under consideration. For instance, the best unimodal *weather* classifier could have been based on the speech recognizer output and not on visual features. The second column BOBO (Best Of Best Of) is the primary run submitted by IBM at TRECVID2003[3]; And it represents the best multimodal model including information fusion across modalities and classifier fusion across different classifiers. For further details on this system submitted by IBM, please refer to TRECVID2003 description[3]. The third column shows results using MaxEnt modeling approach detailed in this thesis. A sample result showing the top 12 retrieved matches for *News Subject Face* is illustrated in Figure 4.4.

From the results it is possible to see that MaxEnt out-performed BOU in 7 concepts and BOBO in 5 concepts. It has to be noted that in the case of BOU and BOBO, the systems had access to a commercial detector and this was used to selectively improve the concept

Concept	BOU	BOBO	FL-21
Outdoors	0.81	0.85	0.98
News Subject Face	0.80	0.73	0.94
People	0.90	0.99	0.92
Building	0.53	0.56	0.55
Road	0.46	0.52	0.67
Vegetation	0.96	0.93	0.91
Animal	0.10	0.10	0.11
Car Truck or Bus	0.68	0.56	0.63
Aircraft	0.38	0.63	0.32
Non Studio Setting	0.97	0.97	0.96
Sports Event	0.81	0.98	0.94
Weather	0.81	0.98	0.68
Mean Precision	0.68	0.73	0.72

Table 4.3: Results of the MaxEnt models compared against the TRECVID2003 benchmark system results. The numbers are Precision at 100 retrieved shots.

detectors[3]. In addition, in the case of BOU, the choice of modality (i.e. audio, text or visual information) and granularity of feature extraction (global versus regional) varied across the different concepts, based on performance on a validation set. In the case of BOBO, the variation spanned not just on input modalities and granularity but also on modality fusion and classifier fusion techniques employed in the final model. On the other hand, the MaxEnt models rely only on the visual features and operate on a fixed feature granularity across all evaluated concepts. The signed t-test between BOU and BOBO indicates significance only at the 90% p-value and the differences between the MaxEnt and BOBO approaches are not statistically significant at this significance value.

4.4 Conclusions

In this chapter the Maximum Entropy approach for automatic semantic annotation of multimedia data with the visual cues were described. This approach was evaluated on the TRECVID2003 corpus and benchmarked against the top ranked systems. The results indicate that this approach is promising and performs as well as the state-of-the-art multimodal systems for automatic semantic annotation despite using a single feature modality. This is a very encouraging result. Further study is needed to evaluate the effect of feature granularity selection (e.g. it could be that concepts such as *weather* and *outdoors* will benefit from global features) and more importantly, inclusion of other modalities (such as audio, closed captions and ASR words which is introduced in the next chapter).

Figure 4.4: *The top 12 results using the MaxEnt models for the News Subject Face concept. Note that the statue is an incorrect classification for this concept.*

Chapter 5

Speech Cues

5.1 Introduction

In the TRECVID contest one tries to estimate the probability of some specific concept given shot. These concepts are often visual, but also different types exist. It is not surprising that in general for visual only concepts visual models work well. However, it could be hard to visual models to detect concept like female speech in confident level. Therefore speech cues are introduced. The TRECVID corpus comprises MPEG videos of broadcast news content. LIMSI provided an automatic speech recognition (ASR) transcript [21] for this corpus. A possible approach to incorporate speech cues would be to use the extracted ASR words directly. Unfortunately this would lead to very sparse feature space that is prone to over fitting. It is possible to reduce the word space dimensionality using unsupervised techniques such as latent semantic indexing (LSI)[26]. The approach presented in this theses leverages advances made in automatic text content extraction and natural language processing.

5.1.1 Features

In this thesis for the experiments the ASR output provided by LIMSI is used as an extracted feature. By using the same ASR output than others in the TRECVID bechmark it makes it easier to compare different approaches to each other and not the performance or behavior of ASR engine.

5.1.2 Entity predicates

Entity predicates are defined to capture correlations between a detected named entity and manual annotation of the training data. These entities are extracted using an Automatic

Content Extraction system [33] (uses MaxEnt also). This system has four main phases (see Figure 5.1). The first phase is to truecase the extracted words. The extracted ASR words are uncased and it has been observed that the case restoration improves the performance of subsequent stages of processing [23]. The computational time spent in true casing is extremely dependent on the average input sentence length. To reduce this time, the ASR text is split into sentences using the following approach. First, the maximum limit for the number of words that the truecaser can handle is set. In the experiments 170 words is used as the limit. A sentence break is inserted at the point of the longest silence in a chunk of words. This process is iterated till all boundaries have fewer than 170 words. Other way to split sentences could be such as histogramming the number of words between silence marks and using a held-out set to choose the minimum silence duration for marking a sentence break. It has to be noted that this did not result in a significant performance change compared to the simple, allthough ad hoc, scheme.

In the next phase, parts-of-speech (POS) information is extracted from the cased sentences. In the third phase, WordNet [9] is used to determine relationships with other words. Finally, the various entities are detected and tagged. Automatic content extraction process results in three distinct entity types: named, nominal and pronominal[32, 33].

$$f_{e,a}(x,y) = \begin{cases} 1 & \text{if } y = a \text{ and } e \in x \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

A predicate of this type is active only if specified entity e is detected in a current shot x and the corresponding manual annotation is a. These predicates capture the entities for current shot. Likewise, predicates based on the entities detected in the previous shot are defined as follows:

$$f_{e_{-1},a}(x,y) = \begin{cases} 1 & \text{if } y = a \text{ and } e_{-1} \in x-1 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

It is also possible to extract these predicates using freely available systems. The POS info can be determined with the OpenNLP tools [4] or with the MXPOST (Maximum Entropy Part of Speech Tagger) [1]. The GATE (General Architecture for Text Engineering) [12] could be used for the named entity detection.

5.1.3 Expanded predicates

These predicates are formed by preparing a list of hypernyms for the current word using WordNet[9]. Then, this list is filtered against the concept vocabulary. For example, the word *wind* will be expanded to weather, weather condition, air current,... . Since *weather news* is present as a concept in the vocabulary, it gets through this filtering process and other

words are discarded. These predicates are presented to model in the same way as the entity predicates.

5.1.4 Activity predicates

The activity predicates are designed to model correlations between class activity and an annotated shot. The activity predicate is active if the class has at least one predicate active. The possible classes include the entity and expanded predicates. In the final model, there are activity predicates for the current, previous and next shot.

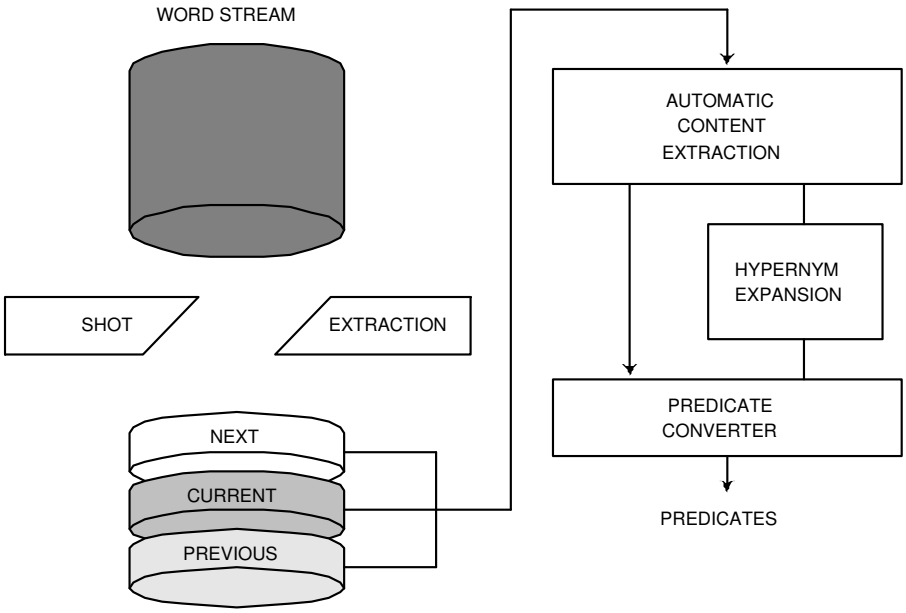

Figure 5.1: *Diagram illustrating the extraction of speech based predicates*

5.2 Experiments on the development set

The Table 5.1 describes experiments conducted.

name	contents
FL-102	Entity and corresponding activity predicates
FL-105	Expanded and corresponding activity predicates
FL-106	Entity, expanded and activity predicates in the same model

Table 5.1: Conducted visual experiments.

Like in the visual part above also here these classifiers were built using the test set and

then evaluated using set concept fusion 2. Specific shot is shown as a positive example of a concept if it is annotated with the concept. Otherwise, shot is treated as a negative example. The Picture 5.2 and Table 5.2 do show the performance on the development set. It has to be noted that the performance increase in the case that the entities (FL-102) and expanded predicates (FL-105) work together in the same model (FL-106) is significant.

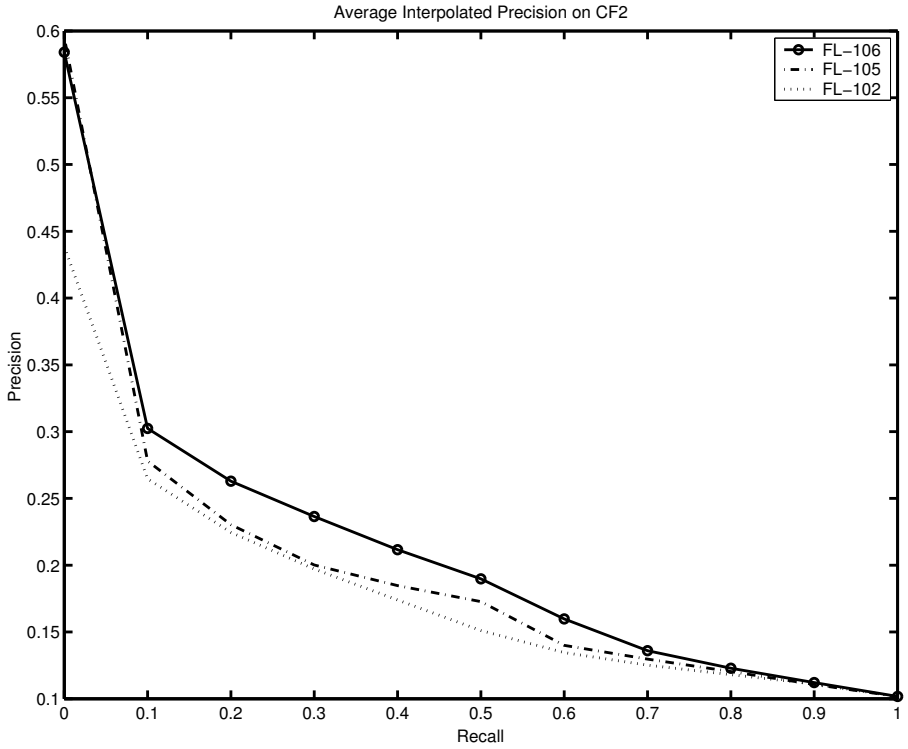

Figure 5.2: *The figure illustrates how performance is affected when more predicate types are added.*

experiment	MP@100	MAP@1000
FL-102	0.110	0.275
FL-105	0.113	0.264
FL-106	0.134	0.293

Table 5.2: The chance in mean average precision and mean precision while number of different predicate type varies.

The final comparison against the best models is carried out in the multimodal part of this thesis. It has to be done due the absence of similar speech only model set.

5.3 Conclusions

In this chapter the Maximum Entropy approach for automatic semantic annotation of multimedia data with the speech cues were described. The experiments on the development set indicate that the best performance is acquired with the setup which uses both entity and expanded predicates in the final model.

Chapter 6

Multimodal Modeling

6.1 Introduction

At the first place it may sound that the best possible classifier for visual concept (e.g. cat) is based on visual cues. And, the best possible classifier for particular speech concept is built top of speech cues. Unfortunately, the case is not that straight forward. First of all it can be hard to predict whether concept is some "modality" only (such as visual). Second, it is well known in literature that multimodal models have attempt to have better performance in general [16, 24]. Lets consider an example. First we train our *cat* model using the best possible visual cues. And, all positive training shots contain positive examples taken from adult cats. Now, model can most likely detect an adult cat from new unseen shots, but how about just born kitten? In the case you have not seen just unborn kitten, my personal opinion is that they look more like small mouses without any hair. This model cannot most likely detect these mouse looking kittens as an adult cats. However, if we had trained our model using multimodal cues (vision + speech) at the first place. The model could know that hearing word "kitten" is as much cat than hearing word "cat" and therefore able to detect kittens as a cats as well. One way to combine visual and speech cues into multimodal setup could start by building separate models for both. After both models are estimated the probabilities for specific concept, this output can be seen as a model output space which have to be mapped into concept probability somehow. For this combining there can be new model which could be trained on another test set to learn when particular model output is correct. Or, mapping can be carried out just by multiplying the probabilities of different models and then normalising them. However, in this thesis both modalities (visual and speech) are proposed to the same model and therefore these kind of methods are not needed.

6.2 Features

The multimodal approach presented here uses the same extracted features as described in the previous sections.

6.3 Predicates

The used predicates are introduced above. The Figure 6.1 illustrates how visual and speech predicates work together together in one multimodal model.

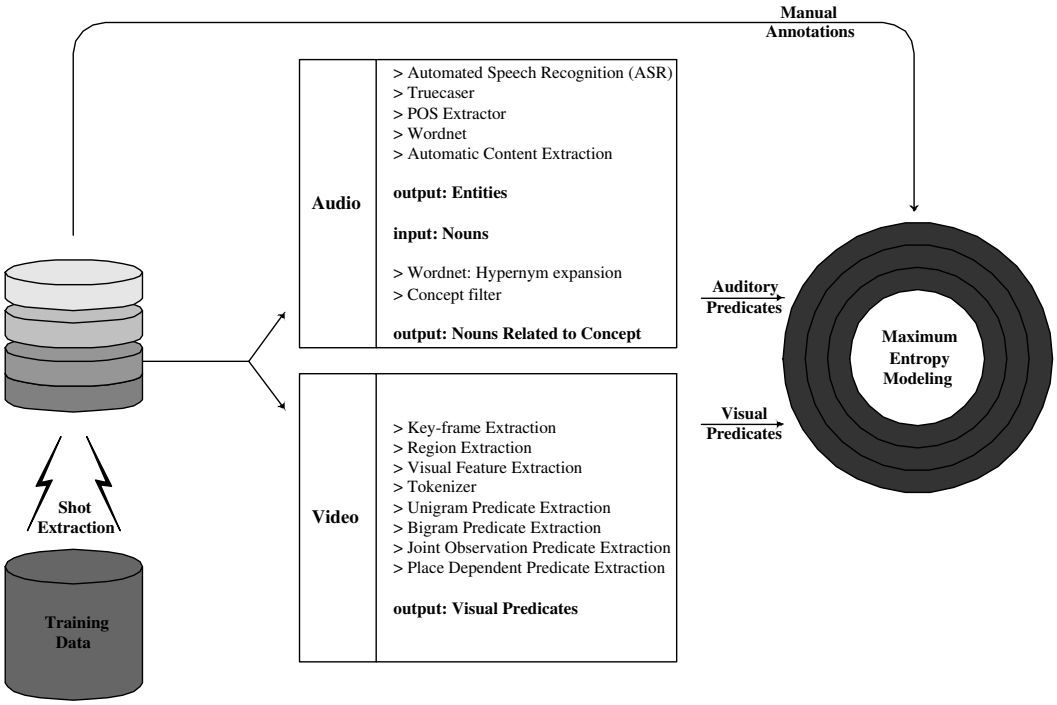

Figure 6.1: *The figure illustrates how multimodal system is built.*

6.4 Experiments on the development set

The Table 6.1 describes experiments conducted. Visual and speech models are described in the previous chapters.

The Table 6.2 and Figure 6.2 illustrates that the performance of multimodal setup is significantly better than any of the best unimodal setups.

name	contents
Speech (FL-106)	Entity, expanded and activity predicates in the same model
Visual (FL-21)	unigram, bigram, place dependent and joint observation predicates
Multimodal (FL-125)	All predicates from the best speech and visual setup in one model

Table 6.1: Conducted multimodal experiments.

experiment	MP@100	MAP@1000
FL-21	0.461	0.252
FL-106	0.134	0.293
FL-125	0.479	0.313

Table 6.2: The change in mean average precision and mean precision while number of different predicate type varies.

6.5 Comparison Against the Best

For the multimodal tests both TRECVID2003 and TRECVID2004 corpora are used. The training of these classifiers is done using the TRECVID2003 corpus. The final performance comparison of these classifiers is carried out with the NIST-evaluated relevance judgments on the TRECVID2004 corpus. For the development partition, NIST has provided ground truth annotations at the video-shot level. In addition, NIST has provided reference key-frames for each shot for the entire corpus and extracted ASR words. The final results for particular classifier is shown in the Table 6.3.

Table 6.3 shows the results of the visual-only, speech-only and multimodal classifiers for each of the 32 concepts in our experiments. The performance is measured using Precision at 100 retrieved shots. The table shows that the Multimodal system has a 6 percent relative improvement compared to the visual-only system. This result is significant at the 95% level, using t-test. The average precision at 100 documents retrieved for visual system is 0.40, for speech 0.24, and for multimodal 0.44. The relative improvement here is 10 percent. The figure 6.3 shows the interpolated Recall-Precision graph for the 3 systems.

The top predicates having either negative or positive influence to the probability of particular concept is shown in Table 6.4. The sentence "... and at the time 1:02 he shoots and scores..." would give good probability to basket scored, but at the same time hearing Clinton or seeing entity WEAPON will cause negative influence.

When compared to IBM models submitted to TRECVID2004, t-test shows that the difference between multimodal system presented in this thesis and IBM-BOM (where a separate classifier was chosen for each individual concept based on the performance on a held-out set) is not statistically significant. The multimodal method presented here for those ten concepts (see Table 6.5) in TRECVID2004 have mean precision 0.21 at 100 documents

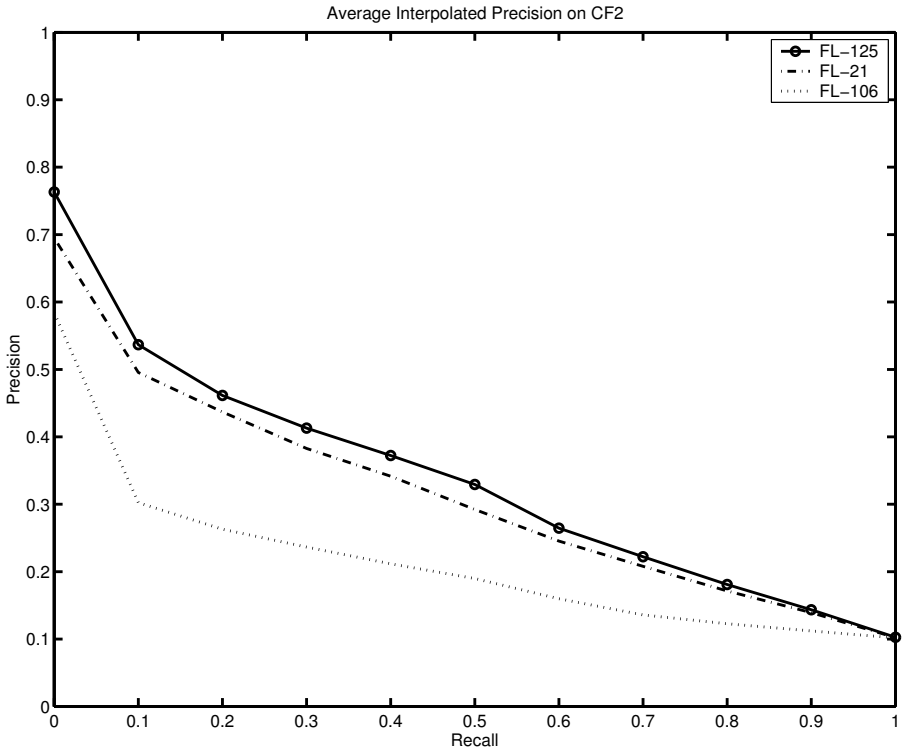

Figure 6.2: *The figure illustrates how performance is affected when moving from unimodal setups ti multimodal.*

retrieved, IBM-BOM has 0.27 and IBM-Mall has 0.30. Difference between the multimodal system and the best system at TRECVID2004 (IBM-Mall) is statistically significant. It has to be noted that IBM-Mall had access to bigger set of extracted features such as closed captions. Also, IBM-Mall is selection from many techniques including SVMs and MaxEnt. The final IBM-Mall model is an ensemble fusion [2] from multiple systems together with post processing such as advertisement filtering to further improve the results. The most important thing is that the visual method presented here was one of the methods in both IBM-Mall and IBM-BOM. This means that those submissions already have possible boost gainable partly from this approach. This, and presented results above make the multimodal approach more interesting.

6.6 Conclusions

This section detailed a general multimodal Maximum Entropy based approach for automatic semantic annotation of multimedia. Multimodal models work significantly better than visual only models. The approach had less features than the state-of-the-art models submitted

Concept	V	S	MM
Airplane takeoff	0.14	0.18	0.20
Animal	0.11	0.15	0.16
Basket scored	0.26	0.11	0.31
Beach	0.15	0.02	0.17
Bill Clinton	0.09	0.17	0.19
Boat/ship	0.11	0.04	0.14
Building	0.43	0.26	0.47
Car	0.54	0.61	0.65
Face	0.97	0.85	0.99
Female Face	0.42	0.43	0.45
Female Speech	0.53	0.31	0.58
Hockey	0.25	0.42	0.06
Indoors	0.88	0.70	0.90
Madeleine Albright	0.00	0.03	0.00
Male Face	0.95	0.59	0.96
Male Speech	0.79	0.73	0.89
Man Made Scene	0.84	0.98	0.83
Nature Non-Vegetation	0.47	0.18	0.51
Nature Vegetation	0.58	0.19	0.59
News Subject Face	0.84	0.39	0.74
News Subject Monologue	0.06	0.13	0.10
Non-Studio Setting	0.90	0.65	0.82
Outdoors	0.81	0.59	0.79
People	0.89	0.42	0.88
People Walk/Run	0.44	0.28	0.53
Physical Violence	0.06	0.05	0.05
Road	0.53	0.29	0.54
Sky	0.98	0.34	0.96
Sporting Event	0.87	0.60	0.93
Studio Setting	0.95	0.73	0.99
Train	0.01	0.01	0.01
Weather News	0.28	0.51	0.41
Mean Precision	**0.50**	**0.37**	**0.53**

Table 6.3: Comparison between visual only (V), speech only (S) and multimodal (MM) setup. The numbers are Precision at 100 retrieved shots.

to TRECVID2004 and yet it performs competitively. This generic, scalable approach shows considerable promise.

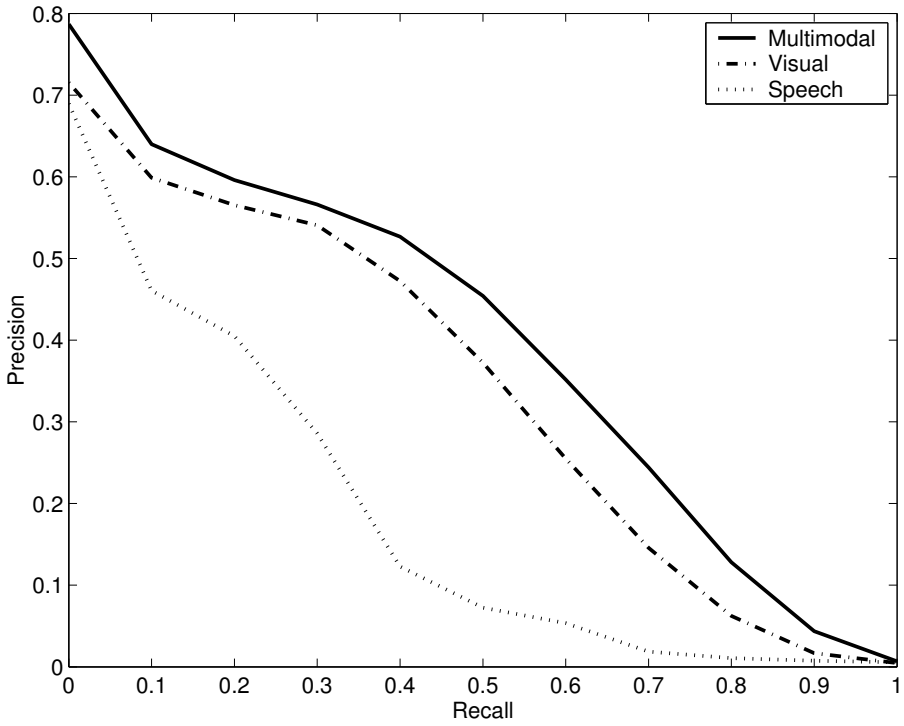

Figure 6.3: *The figure shows the interpolated recall-precision graph for the 2 unimodal systems and the multimodal system.*

Figure 6.4: *The figure shows key frames from the first 8 shots retrieved for the concept car.*

Data	Type	Location	Influence
made	expanded	previous	+
splash	expanded	current	+
made	expanded	current	+
players	expanded	current	+
TIME	entity	current	+
DATE	entity	current	+
basketball	expanded	previous	+
basketball	expanded	current	+
EVENT SPORTS	entity	previous	+
shot	expanded	current	+
FOOD	entity	current	-
WEAPON	entity	current	-
fencing	expanded	previous	-
skating	expanded	previous	-
skating	expanded	current	-
GEOLOGICALOBJ	entity	previous	-
FOOD	entity	previous	-
ANIMAL	entity	previous	-
clinton	expanded	current	-
landing	expanded	current	-

Table 6.4: Some predicates having either big negative or positive influence to the probability of the concept basket scored.

Concept	MM	IBM-BOM	IBM-Mall
Boat/ship	0.14	0.24	0.23
Madeleine Albright	0.00	0.08	0.08
Bill Clinton	0.19	0.28	0.32
Train	0.01	0.01	0.00
Beach	0.17	0.10	0.14
Basked scored	0.31	0.45	0.46
Airplane takeoff	0.20	0.15	0.15
People Walk/Run	0.53	0.83	0.84
Physical violence	0.05	0.17	0.30
Road	0.54	0.42	0.51
Mean Precision	**0.21**	**0.27**	**0.30**

Table 6.5: Results of the MaxEnt models compared against the TRECVID2004 benchmark system results. The numbers are Precision at 100 retrieved shots.

Chapter 7

Conclusions and Future Work

This thesis details three general Maximum Entropy based approaches for automatic annotation of multimedia. It was shown that multimodal approach works significantly better than any of the unimodal ones. The performance of final models are at the same level as state-of-the-art systems submitted to TRECVID. This gives approach big promise because the winning approach in the TRECVID contest were built with additional information in their use (such as additional features and commercial detector, see sections above).

Presented approach (like all approaches) just try to model real world. And, the methods used today are still far from the human vision system. There are several place we could improve. One of the main problems when acquiring real general approach is the generalization. The system developed using some set(s) for a particular domain (like here for news) do not usually generalize to a different domain [37]. But, these system may are still usable in the domain it self. And, the presented method in this thesis is sufficiently general that it can easily be used in other domains as well (e.g. face recognition task for identifying faces for multimodal speech detection [10]. There are also different type of generalization. Generalization over different languages. The visual part of the presented method will work as-is for any language. But, the speech part will need some language update (entity, noun detection,...). Fortunately this kind of tasks are already going [33].

One of the missing features of presented approach is experience. A human (hopefully) learns to recognize how the road looks like when going to school or work. However, presented approach will forget shown shot and made decision right away we present new shot to it. This leads to other problem as well, to adaptation. If the system cannot learn from made decisions it won't automatically adapt to new environment. Which would not be that hard if the environment is changing slowly.

The speech part gives an experimenter easy access to see what MaxEnt model has basically learnt and based on what it is going to do further decisions (see speech part). Visual

side does not have such easy way for this. The one aspect of future work will be in finding such methods that allow the visualisation of learnt visual model. In this way - after training - we could ask model to present us e.g. a general picture presenting the concept.

One of the future interests is to try more modalities such as closed captions. And, see whether adding more modalities still makes improvements to performance. It may also be that the count of presented predicates to the model is already too high when compared to positive training examples. If so, then adding new predicates would easily lead to overfiffing with the training data. This means that model finds some meaningless regularity which does not present the actual concept [35].

It would make also more sense if we first estimate the probability of face and then figure out if it is female or male face. Currently all classifiers give their output given the same input to all. In general this would mean that we should test stronger hypothesis first and weaker ones then. This would most likely reduce computational complexity as well.

In addition, multiple binary classifiers were built and used in these experiments. A natural question to address is the performance difference between a single multi-way MaxEnt classifier for all concepts versus multiple binary classifiers. In addition, it has to be noted that the ground truth annotations can be quite variable in quality; Frequently, the common objects are not marked. For instance, concepts such as *Outdoors* tend to be missing in many annotations despite being present in the shot. This needs to be explicitly accounted for by allowing the model to handle *unlabeled* objects and regions in a video shot.

Bibliography

[1] Adwait Ratnaparkhi. Maximum Entropy Part-Of-Speech Tagger. *In Proceedings of the Empirical Methods in Natural Language Processing Conference*, May 1996.

[2] A. Amir, J. Argillander, M. Berg, S. Chang, W. Hsu, G. Iyengar, and et al. IBM research TRECVID2004 video retrieval system. In *Proceedings of the TRECVID2004 Conference*. NIST, 2004.

[3] A. Amir, M. Berg, S. Chang, W. Hsu, G. Iyengar, and et al. IBM research TRECVID2003 video retrieval system. In *Proceedings of the TRECVID2003 Conference*. NIST, 2003.

[4] Jason Baldridge, Tom Morton, and Gann Bierner. openNLP tools api. http://opennlp.sourceforge.net/, version 2.2.0, 2004.

[5] Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[6] R. Brunelli, O. Mich, and C. Modena. A survey on video indexing, 1996.

[7] J.N. Darroch and D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[8] M. Flickner, H. Sawhney, and et al. Query by image and video content: The qbic system. *IEEE Computer*, 28(9):23–32, 1995.

[9] G. A. Miller. WordNet: A Lexical Database. *Communications of the ACM*, 33(11):39–41, 1995.

[10] G. Potamianos, C. Neti, G. Gravier, and A. Garg. Automatic Recognition of audio-visual speech: Recent progress and challenges. *Proceedings of the IEEE*, 91(9), September 2003.

[11] E. Bruce Goldstein. *Sensation and Perception*. Wadsworth, 2002.

[12] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, July 2002.

[13] I. J. Good. Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contigency Tables. *The Annals of Mathematical Statistics*, 34:911–934, 1963.

[14] IBM. Homepage of CueVideo, July 2004. http://www.almaden.ibm.com/projects/cuevideo.shtml.

[15] International Organisation for Standartisation. Coding of Moving Pictures and Audio: MPEG-7 Overview, March 2003. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[16] G. Iyengar, H. J. Nock, and C. Neti. Discriminative model fusion for semantic concept detection and annotation in video. In *Proceedings of ACM Multimedia Conference*, Berkeley, CA, November 2003. ACM.

[17] J. Argillander, G. Iyengar and Harriet Nock. Semantic Annotation of Multimedia Using Maximum Entropy Models. *ICASSP*, 2005.

[18] E. T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106:620–630, 1957.

[19] E. T. Jaynes. Notes on present status an future prospects. *Maximum Entropy and Bayesian Methods*, pages 1–13, 1991.

[20] Jiwoon Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *Image and Video Retrieval: Third International Conference, (CIVR)*, Lecture Notes in Computer Science, Dublin, Ireland, July 2004. Springer-Verlag.

[21] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. In *Speech Communications*, pages 89–108, 2002.

[22] V. Lavrenko, S. L. Feng, and R. Manmatha. Statistical models for automatic video annotation and retrieval. In *Proceedings of ICASSP*, Montreal, Canada, May 2004. IEEE.

[23] L.V. Lita, A. Ittycheriah, S. Roukos, N. Kambhatla. tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159, July 2003.

[24] M-Y. Chen, A. Hauptmann. Multi-modal Classification in Digital News Libraries. In *Joint Conference on Digital Libraries (JCDL'04)*, June 2004.

[25] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[26] M.W. Berry, S.T. Dumais, and T.A. Letsche. Computational Methods for Intelligent Information Access. In *Proceedings of Supercomputing'95*, San Diego, CA, December 1995.

[27] Milind R. Naphade, Igor Kozintsev, Thomas S. Huang, and Kannan Ramchandran. A factor graph framework for semantic indexing and retrieval in video. In *CBAIVL '00: Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00)*, page 35. IEEE Computer Society, 2000.

[28] NIST. *TREC Video Retrieval Evaluation Conference(TRECVID2003)*, Gaithersburg, MD, November 2003.

[29] NIST. TREC Measures, November 2004. http://trec.nist.gov/pubs/trec10/appendices/measures.pdf.

[30] NIST. *TREC Video Retrieval Evaluation Conference(TRECVID2004)*, Gaithersburg, MD, November 2004.

[31] J.F.G. de Freitas P. Duygulu, K. Barnard and D.A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conf. on Computer Vision*, 2002.

[32] H. Jing R. Florian, A. Ittycheriah and T. Zhang. Named entity recognition through classifier combination. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada, 2003.

[33] R. Floridian, H. Hassan, A. Ittycheriah, H. Jing and et al. A Statistical Model for Multilingual Entity Detection and Tracking. *Proceedings of the NAACL/HLT*, 2004.

[34] A. Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, Philadelphia, PA, 1998.

[35] Stuart Russel and Peter Norvig. *Artificial Inteligence*. Prentice Hall, 2003.

[36] John R. Smith and Shih-Fu Chang. VisualSEEk: A fully automated content-based image query system. In *Proceedings of the ACM Multimedia Conference*, pages 87–98. ACM, 1996.

[37] Thomas M. Strat. *Natural object recognition*. Springer-Verlag New York, Inc., 1992.

[38] Martin Szummer and Rosalind W. Picard. Indoor-outdoor image classification. In *International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, pages 42–51. IEEE, 1998.

[39] T. Utsuro, T. Miyata, and Y. Matsumoto. Maximum entropy model learning of subcategorization preference. In *Proceedings of the 5th Workshop on VLC*, pages 246–260, 1997.

Chapter 8

Appendix A

The paper "Semantic Annotation of Multimedia using Maximum Entropy Models", published in ICASSP 2005, is attached.

SEMANTIC ANNOTATION OF MULTIMEDIA USING MAXIMUM ENTROPY MODELS

Janne Argillander, Giridharan Iyengar, Harriet Nock

IBM TJ Watson Research Center
Yorktown Heights, NY 10598
Email: janne@fi.ibm.com,{giyengar,hnock}@us.ibm.com

ABSTRACT

In this paper we propose a Maximum Entropy based approach for automatic annotation of multimedia content. In our approach, we explicitly model the spatial-location of the low-level features by means of specially designed predicates. In addition, the interaction between the low-level features is modeled using joint observation predicates. We evaluate the performance of semantic concept classifiers built using this approach on the TRECVID2003 corpus. Experiments indicate that our model performance is on par with the best results reported to-date on this dataset; Despite using only unimodal features and a single approach towards model-building. This compares favorably with the state-of-the-art systems which use multimodal features and classifier fusion to achieve similar results on this corpus.

1. INTRODUCTION

Growing amounts of multimedia content, especially video have reached a critical point where methods for indexing, searching, and efficient retrieval are expressly needed to manage the information load. The amount of multimedia content that is already present in most consumer hard-drives makes manual annotation (and consequently, indexing using high-level keywords) impossible. There has been some effort in using query-by-example (QBE) to seek into multimedia content (e.g [1, 2] amongst many others). While QBE is a powerful paradigm, its reliance on low-level perceptual properties is counter to the *semantic* nature of most user queries[1, 3]. Query-by-keyword (QBK) systems, where the user queries the content using semantic descriptors, are getting more and more attention. These systems typically require at least two processing stages: A primary *training* phase where the system is taught to identify specific concepts from a pre-defined vocabulary; A secondary *annotation* phase where the system (semi-)automatically annotates previously unseen content with these newly learned concepts.

In this paper, we focus on our recent work in using Maximum Entropy (MaxEnt) modeling techniques for automatic annotation of multimedia content. In our particular approach, the problem is formulated similar to a multiple instance learning problem. By this we mean that the annotations are specified at the level of the entire image or video shot. That is, given an annotation such as *face*, we know that the particular training shot contains a face but not its precise location within the shot. This is aimed at reducing the *acquisition effort* involved in training these semantic concept models[1].

[1]See our previous discussion on the three distinct dimensions along which we believe concept modeling systems need to be measured[4].

The rest of the paper is organized as follows: In section 2 we provide a quick summary of related work in this area. The details of our MaxEnt modeling approach follows in section 3. The dataset and experiments are detailed in section 4 followed by conclusions.

2. RELATED WORK

There is extensive literature in object detection (especially human faces) where the extent of an object is well-marked in an image. There is relatively limited literature in automatic image annotation where the physical extent of objects are not specified. In one set of approaches, techniques from statistical machine translation were applied to the problem of image annotation. In these approaches it is assumed that the annotation and the associated image are translations of each other and with a suitable of *tokenization* of the image features, standard machine translation models have been applied with some success[5]. Motivated from a cross-lingual information retrieval perspective, Lavrenko et al.[6] approach image annotation as an example-based learning problem where perceptual similarity in the image space is assumed to generate similar annotation words. Both these approaches have been demonstrated on relatively small datasets (5000 images from COREL dataset) and they remain to be evaluated in larger contexts such as what is attempted in this paper (e.g. 80000 shots from TRECVID2003 corpus[7]). Motivated by the under-constrained nature of the annotation problem together with the non-independent nature of low-level image features, we approach this in a Maximum Entropy setting which has had remarkable success in many Natural Language Processing tasks such as sentence-boundary detection and parts-of-speech tagging[8, 9]. A similar approach using MaxEnt for image annotation was proposed in[10]. The novelties of our approach are two-fold: We model the spatial- and joint-dependence between low-level features using specially designed predicates. We believe such information is important for objects that have a well-defined spatial composition (e.g. faces). In addition, we evaluate our approach on a much larger corpus (TRECVID2003). Furthermore, we present a comparison of our approach with previously published results on the TRECVID2003 concept detection task[7, 11].

3. MAXIMUM ENTROPY APPROACH FOR MULTIMEDIA ANNOTATION

In MaxEnt modeling, we assume that a random process produces an output (label) y given a context x. In multimedia annotation, y, which is a member of a finite set (vocabulary) Y, can be seen as a label for a specific shot. And x, a member of a finite set X,

as extracted information (features) from the current frame. Training data is presented in pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. The task is to learn possible correlations between x and y, and to build statistical models that can be used to annotate previously unseen shots automatically. The empirical probability distribution function (pdf) based on training data is as follows

$$\tilde{p}(x, y) = \frac{1}{n}\text{freq}(x, y) \tag{1}$$

Where *freq* is the count of a specific pair (x, y) in the training data. In real world applications, the training set size is finite. Therefore, the empirical distribution is a poor estimate of the joint pdf. Based on this partial information, MaxEnt modeling can be used to estimate the pdf that generated the empirical distribution $\tilde{p}(x, y)$ in an unbiased way[12]. At the core of the modeling process are feature functions. In this paper we prefer the term *predicates* over *feature functions* to avoid confusion with extracted low-level image features. These predicates are used to specify constraints on the model. In MaxEnt, the process of defining predicates is central to modeling: The goodness of the models is dependent on the ability of these predicates to capture relevant information. We now detail the different predicates used to capture a variety of spatial and co-occurrence properties of the low-level image features. We note again that this is a fundamental difference between our approach for multimedia annotation over previous work[10].

In our experiments, we extract 3 types of low-level image features from each video shot: Lab space color moments (mean, variance, skewness and kurtosis for each channel), Edge orientation histogram (Edge strength and orientation values at each pixel, each quantized to 8 bins) and summary statistics of grey-level co-occurrence matrices (entropy, energy and contrast values). Together, these form our 3 different low-level descriptors which we will term *Color*, *Edge* and *Texture* in further discussions. Furthermore, we partition each shot key-frame (comprising 350×240 pixels) into 35 regions (50×48 pixels each) and extract the feature descriptors for each of these 35 regions.

3.1. Unigram predicates

Unigram predicates are defined to capture the co-occurrence statistics between a specific tokenized descriptor and manual annotation of the training data. All unigram predicates used in this paper have following form:

$$f_{cd^i,a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i \in x^i, i = 0, 1, 2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

A predicate of this type is active only if tokenized descriptor cd is in current frame x and the corresponding manual annotation is a. The total number of unique unigram predicates in our model is (descriptor count x cluster size) 3x25=75.

3.2. Place Dependent Unigram Predicates

Place dependent unigram predicates are designed to capture location specific statistics. For instance, these predicates help the model learn that regions corresponding to *sky* are usually in upper parts of a key-frame.

$$f_{cd^i,a}(x^i_r, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i = x^i_r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Where region r has values $0..34$ and descriptor i takes values $0, 1, 2$. The place dependent predicate is active only if the tokenized descriptor cd is in region r of the current frame which has the annotation a. The total number of such predicates in our model is (descriptor count x region count x cluster size) 3x35x25=2625.

3.3. Bigram Predicates

In our work we have experimented with two types of bigram predicates:horizontal and vertical. These predicates model the relationship between neighboring regions. Below is an example horizontal bigram predicate which is active only if tokenized descriptor cd_r and its horizontal neighbor cd_{r+1} is adjacent in current frame x with annotation a.

$$f_{cd^i_r+cd^i_{r+1},a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i_r + cd^i_{r+1} \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where the region r take values so that the adjacent region on right cd^i_{r+1} is in the row. The following equation illustrates a vertical bigram predicate which is active only if tokenized descriptor cd_r and its vertical neighbor cd_{r+7} are also adjacent in current frame x.

$$f_{cd^i_r+cd^i_{r+7},a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i_r + cd^i_{r+7} \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Where the region r take values so that the adjacent region below is in same column $r = 0..27$.

Both types of bigrams are constructed by combining the tokenized features in the product space of the unigram predicates. This choice imposes the possibility of obtaining bigram values that are not supported in the training data, resulting primarily from the sparseness of the product space. To counter this, we employ an approach inspired from class-based language models in speech processing. When two unigrams are composed into a bigram, we treat them differently. We start with few clusters for the composed unigrams and slowly increase the number of clusters such that the number of unique bigram predicates observed (in the training data) at each step matches the total possible bigram product space values. We stop at the largest cluster size for which this condition is met in the training data.

3.4. Joint Observation Predicates

The predicates discussed so far model individual low-level feature descriptors (i.e. Color, Edge, Texture). We now illustrate predicates that model the interactions between the various low-level feature descriptors.

$$f_{cd,a}(x, y) = \begin{cases} 1 & \text{if } \forall i \; y = a \text{ and } cd^i \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This joint observation predicate is active only if all low-level descriptors are present in a given region. In our experiments we work with 144 such joint observation predicates, chosen using validation data. Figure 1 illustrates the various types of predicates used in our model.

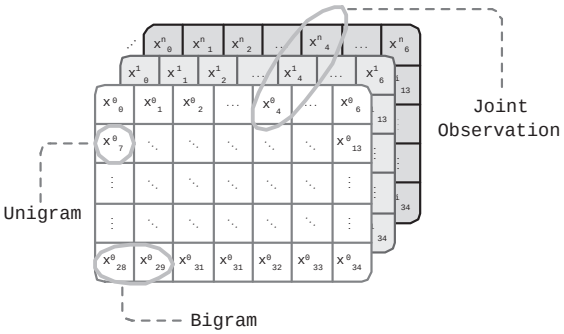

Fig. 1. *The figure illustrates the 35-region grid partition of shot key-frames. Shown in the figure are the place-dependent unigrams, horizontal bigrams and joint-observation predicates.*

3.5. Model Preparation

In this work, we built a distinct binary classifier for each semantic concept that we used to evaluate the models. Keeping with the Multiple Instance learning approach, each labeled example containing the target concept annotation[2] and is considered a positive example for those classifiers. All training instances that do not contain the target annotation are marked as negative examples. These empirical expectations of the various predicates provide constraints for the MaxEnt modeling. We use the open source MaxEnt modeling toolkit[13] in our experiments and in particular use the Generalized Iterative Scaling (GIS) algorithm[14] with smoothing to solve for the conditional probability density function.

3.6. Automatic Annotation of Unseen Multimedia Content

Semantic concept annotation of unseen multimedia content proceeds in the following manner. First, the low-level feature descriptors are extracted from the data, using the same 35-grid partitioning of the shot key-frames. For each concept to be predicted, the set of all active predicates relevant to the concept are extracted from the feature descriptors. We now have enough information to estimate the conditional probability of the particular annotation for the shot key-frame.

4. DATASET AND EXPERIMENTS

We now detail the dataset used to evaluate the MaxEnt models for semantic concept annotation of multimedia data and present our experimental results.

4.1. Dataset and model preparation

We use the TRECVID2003 corpus comprising 120 hours of broadcast news videos for our experiments. This corpus is further divided into approximately evenly sized test and development partitions. We compare the performance of our system with the NIST-evaluated relevance judgments reported on the test partition. For the development partition NIST has provided ground truth annotations at the video-shot level. In addition, NIST has provided reference key-frames for each shot for the entire corpus. For each of these reference key-frames we extract the specified low-level feature descriptors on a 35-grid layout as indicated earlier in the paper. We associate the shot-level ground truth annotations to each of the reference key-frames in development partition. We note that these annotations are provided at the shot-level and do not specify spatial or temporal boundaries of objects within a shot (i.e. we know that a face appeared in the shot but we do not know *when* and *where* within this shot). We selected 12 of the 17 benchmarked concepts from TRECVID2003; We removed audio concepts (*Female Speech*), abstract concepts (*Physical Violence*), specific person concepts (*Madeleine Albright*), camera operations (*Zoom-in*) and multimodal concepts (*News Subject Monologue*). In the case of audio and multimodal concepts, these were removed because our low-level features do not capture relevant information. Camera operations do not belong in the same category of concepts as the rest of the concepts. Both *Physical Violence* and *Madeleine Albright* had very few training examples. We note that the performance of the benchmark systems on these concepts were quite low as well. For each of the selected concepts, we build a MaxEnt classifier as previously stated. These trained classifiers are then used to annotate the test corpus.

4.2. Results

NIST provides pooled relevance judgments and since our system was not part of this pooling, it would be unfairly biased to compare our system with the pooled judgments. To make comparisons valid, we choose to evaluate the different systems using precision at the top 100 retrieved shots as opposed to the average precision metric that is used by NIST. Furthermore, we restrict our comparison to two of the top 10 semantic concept detection systems at TRECVID2003: the best performing (multimodal) system and the best unimodal system[11]. All results are detailed in Table 1. The table also details the 12 concepts classifiers that we built. The first column BOU (Best Of Unimodal) is formed from the set of models by selecting the best performing unimodal classifier for the semantic concept under consideration. For instance, the best unimodal *weather* classifier could have been based on the speech recognizer output and not on visual features. The second column BOBO (Best Of Best Of) is the primary run submitted by IBM at TRECVID2003[11]; And it represents the best multimodal model including information fusion across modalities and classifier fusion across different classifiers. For further details on this system, please refer to our TRECVID2003 description[11]. The third column shows results using MaxEnt modeling approach detailed in this paper. A sample result showing the top 12 retrieved matches for *News Subject Face* is illustrated in Figure 2.

From the results we see that MaxEnt out-performed BOU in 7 concepts and BOBO in 5 concepts. We note here that in the case of BOU and BOBO, the systems had access to a commercial detector and this was used to selectively improve the concept detectors[11]. In addition, in the case of BOU, the choice of modality (i.e. audio, text or visual information) and granularity of feature extraction (global versus regional) varied across the different concepts, based on performance on a validation set. In the case of BOBO, the variation spanned not just on input modalities and granularity but also on modality fusion and classifier fusion techniques employed in the final model. On the other hand, the MaxEnt models

[2]We note here that each training example has multiple concept labels in its ground truth annotation. E.g. A shot may be annotated as a face, outdoors, sky etc.

Concept	BOU	BOBO	MaxEnt
Outdoors	0.81	0.85	0.98
News Subject Face	0.80	0.73	0.94
People	0.90	0.99	0.92
Building	0.53	0.56	0.55
Road	0.46	0.52	0.67
Vegetation	0.96	0.93	0.91
Animal	0.10	0.10	0.11
Car Truck or Bus	0.68	0.56	0.63
Aircraft	0.38	0.63	0.32
Non Studio Setting	0.97	0.97	0.96
Sports Event	0.81	0.98	0.94
Weather	0.81	0.98	0.68
Mean Precision	**0.68**	**0.73**	**0.72**

Table 1. Results of the MaxEnt models compared against the TRECVID2003 benchmark system results. The numbers are Precision at 100 retrieved shots.

Fig. 2. *The top 12 results using the MaxEnt models for the News Subject Face concept. Note that the statue is an incorrect classification for this concept.*

rely only on the visual features and operate on a fixed feature granularity across all evaluated concepts. We further note that a signed t-test between BOU and BOBO indicates significance only at the 90% confidence level and the differences between the MaxEnt and BOBO approaches are not statistically significant.

5. CONCLUSIONS

In this paper we detailed a Maximum Entropy approach for automatic semantic annotation of multimedia data. This approach was evaluated on the TRECVID2003 corpus and benchmarked against the top ranked systems. The results indicate that this approach is promising and performs as well as the state-of-the-art multimodal systems for automatic semantic annotation despite using a single feature modality. This is a very encouraging result. Further study is needed to evaluate the effect of feature granularity selection (e.g. we posit that concepts such as *weather* and *outdoors* will benefit from global features) and more importantly, inclusion of other modalities (such as audio and speech. E.g. *weather* has a distinct vocabulary) on the performance of the MaxEnt models.

In addition, we have built multiple binary classifiers in these experiments. A natural question to address is the performance difference between a single multi-way MaxEnt classifier for all concepts versus multiple binary classifiers. In addition, we note that the ground truth annotations can be quite variable in quality; Frequently, the common objects are not marked. For instance, concepts such as *Outdoors* tend to be missing in many annotations despite being present in the shot. This needs to be explicitly accounted for by allowing the model to handle *unlabeled* objects and regions in a video shot. We intend to address these issues in a future paper.

6. REFERENCES

[1] John R. Smith and Shih-Fu Chang, "VisualSEEk: A fully automated content-based image query system," in *Proceedings of the ACM Multimedia Conference*. ACM, 1996, pp. 87–98.

[2] M. Flickner, H. Sawhney, and et al, "Query by image and video content: The qbic system," *IEEE Computer*, vol. 28(9), pp. 23–32, 1995.

[3] Martin Szummer and Rosalind W. Picard, "Indoor-outdoor image classification," in *International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*. IEEE, 1998, pp. 42–51.

[4] G. Iyengar, H. J. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proceedings of ACM Multimedia Conference*, Berkeley, CA, November 2003, ACM.

[5] J.F.G. de Freitas P. Duygulu, K. Barnard and D.A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Seventh European Conf. on Computer Vision*, 2002.

[6] V. Lavrenko, S. L. Feng, and R. Manmatha, "Statistical models for automatic video annotation and retrieval," in *Proceedings of ICASSP*, Montreal, Canada, May 2004, IEEE.

[7] NIST, *TREC Video Retrieval Evaluation Conference(TRECVID2003)*, Gaithersburg, MD, November 2003.

[8] A. Ratnaparkhi, *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1998.

[9] T. Utsuro, T. Miyata, and Y. Matsumoto, "Maximum entropy model learning of subcategorization preference," in *Proceedings of the 5th Workshop on VLC*, 1997, pp. 246–260.

[10] Jiwoon Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation," in *Image and Video Retrieval: Third International Conference, (CIVR)*, Dublin, Ireland, July 2004, Lecture Notes in Computer Science, Springer-Verlag.

[11] A. Amir, M. Berg, S. Chang, W. Hsu, G. Iyengar, and et al, "IBM research TRECVID2003 video retrieval system," in *Proceedings of the TRECVID2003 Conference*. NIST, 2003.

[12] E. T. Jaynes, "Information theory and statistical mechanics," *The Physical Review*, vol. 106, pp. 620–630, 1957.

[13] Jason Baldridge, Tom Morton, and Gann Bierner, "openNLP maximum entropy modeling toolkit," http://maxent.sourceforge.net/, version 2.2.0, 2004.

[14] J.N. Darroch and D. Ratcliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.