INEXPENSIVE FUSION METHODS FOR ENHANCING FEATURE DETECTION

Peter Wilkins, Tomasz Adamek, Noel E. O'Connor and Alan F. Smeaton

Centre for Digital Video Processing & Adaptive Information Cluster

Dublin City University, Dublin

Alan.Smeaton@computing.dcu.ie

ABSTRACT

Recent successful approaches to high-level feature detection in image and video data have treated the problem as a pattern classification task. These typically leverage the techniques learned from statistical Machine Learning, coupled with ensemble architectures that create multiple feature detection models. Once created, co-occurrence between learned features can be captured to further boost performance. At multiple stages throughout these frameworks, various pieces of evidence can be fused together in order to boost performance. These approaches whilst very successful are computationally expensive, and depending on the task, require the use of significant computational resources. In this paper we propose two fusion methods that aim to combine the output of an initial basic statistical machine learning approach with a lower-quality information source, in order to gain diversity in the classified results whilst requiring only modest computing resources. Our approaches, validated experimentally on TRECVid data, are

designed to be complementary to existing frameworks and can be regarded as possible replacements for the more computationally expensive combination strategies used elsewhere.

1. INTRODUCTION

The purpose of a video information retrieval system is to satisfy an information need for a given user by returning relevant ‘chunks’ of video. In order for a retrieval system to perform this function it needs to first index the video to be searched and the basic sources of data for video indexing can be divided into audio and visual data.

Audio data is regularly exploited in video information retrieval, most frequently through the use of Automatic Speech Recognition (ASR) techniques which produce text transcripts of the speech. Once this text data is available, traditional text-based information retrieval can be employed to aid in the retrieval process. Because of the relatively poor accuracy achievable by ASR, retrieval systems that use speech data are highly effective when the corpus of video is speech-heavy such as in broadcast TV news or documentaries for example, and is self descriptive of the visual content [12]. However if the speech needs to be translated into the language of the user, or if the video is speech-light such as in rushes for example, then an over-reliance on speech data can adversely effect retrieval performance [4].

Visual data is much harder to index and interpret than its audio counterpart. The basic level of visual indexing is to utilize what are known as ‘low-level’ features to represent the visual data of an image [25]. These often take the form of describing the colour or edge distributions of an image. Whilst this data is useful for content-based search where a collection can be ranked based on the similarity to an example query image, by itself it provides no semantic information about what is being retrieved. This problem is referred to as the ‘semantic gap’ [25]. To address this problem techniques are used from computer

vision to classify images based on semantic labels.

For instance, if we successfully classify our collection based on ‘concepts’ such as ‘car’, ‘cityscape’ and ‘persons’, then the user can issue a query such as “retrieve videos containing cars in a city setting with people”. This allows for far greater semantic freedom in expressing retrieval queries in the visual domain, and is generally seen as the direction that video retrieval systems should be looking to exploit [8]. We’ll refer to this problem of classifying images according to a set of semantic labels as *High-Level Feature Detection*.

Whilst it is possible for retrieval systems to create single specialized classifiers for particular semantic concepts, this has limited applicability in any robust retrieval system as it is not possible to create a specific classifier for every concept that might be required. Generic approaches to High-Level Feature Detection are now utilized by video retrieval systems, as these are not constrained to any particular concept [22]. These generic approaches (highlighted in Section 2) make use of machine learning techniques and achieve good levels of performance. Machine learning, particularly approaches which utilize Support Vector Machines (SVM) for generic classification are computationally expensive once multiple low-level features are used in different permutations, the outputs of which are then used as inputs into further multi-tiered classification tasks. With each successive model to be trained, a search of the parameter space needs to be performed (via cross-validation) in order to optimize performance. A major outcome of these systems is the need to continually fuse together various combinations of data. Furthermore, these systems make use of significant computing resources in order to handle the large amount of processing that is required [7].

Our approach to High-Level Feature Detection presented in this paper utilizes a basic machine learning baseline onto which we add additional sources of information in order to achieve performance

gain. We will describe our baseline Support Vector Machine (SVM) output, which whilst relatively fast achieves good performance in classification tasks. We then introduce two methods of further evidence combination which seek to improve upon the baseline. Whilst we will highlight where these further methods improved or reduced the accuracy of the initial classification, we view these approaches as complementary to existing approaches, rather than as a replacement, particularly in the area of combinations of evidence. These approaches were first presented in our TRECVid workshop submission [30], and later expanded upon in our CBMI 2007 conference submission [29].

The rest of this paper is organized as follows. The remainder of this section will briefly discuss the TRECVid evaluation campaign, which provided the context for this work and will specify terminology used throughout the rest of the paper. Section 2 will examine other work on generic High-Level Feature Detection. Section 3 will detail our baseline SVM output, and our two combination algorithms for enhancing detection. Next in Section 4 we analyze where each approach succeeded and failed, wrapping up with conclusions in Section 5.

1.1. TRECVid

All of the work presented in this paper was conducted during the 2006 TRECVid [22, 13] evaluation campaign. TRECVid, formally known as the TREC Video Retrieval Evaluation, is organized by the National Institute of Standards and Technology (NIST) in the US. The goal of TRECVid is "to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results" [1]. More specifically TRECVid promotes research in content-based retrieval of digital video and compares differing approaches by utilizing open, metrics-based evaluation. TRECVid is an evaluation campaign that has truly global reach with 54 partic-

ipants from six continents actively participating in the 2006 campaign. Running since 2001, TRECVid has continued to grow in size, both in terms of number of participants as well as the amount of digital video content utilized for benchmarking and now has a significant impact on research into content-based video analysis and retrieval.

For video retrieval evaluation to take place, a specification of a common retrieval unit is required. For TRECVid the common retrieval unit is a 'shot'. A shot is a segment of video that is at least 2 seconds long, and is bookended by cuts. To represent a 'shot', NIST provides 'keyframes' which are images taken from the shot. These are of two types, Representative Keyframes (RKF) which is a single image to represent the shot as a whole, and Non-Representative Keyframes (NRKF) which represent multiple parts of a shot if the visual contents within a shot varies. In the collection there will always be one RKF per shot, and 0 or more NRKF images for the same shot.

The TRECVid 2006 campaign provided the largest amount of video data of all previous TRECVids with 160 hours of video test data to be used by participants. The data used for 2006 is broadcast TV news data, that is, production video from TV news broadcasts. The data was a multi-lingual corpus comprising Arabic, Chinese and English sources. Participants in TRECVid 2006 were provided with 160 hours of MPEG-1 digital video, ASR transcripts, Machine Translation (MT) output of non-English audio, common shot boundaries and still images from each shot (known as 'keyframes').

TRECVid offers four tasks in which research groups can participate. Of these, this paper concentrates on the second task, that of High Level Feature Detection. The High-Level Feature Detection task for 2006 was given the test collection and a set of 39 semantic features to detect, find for each of those features the top 2000 shots which contain that feature. Participants were also provided with annotated training data which came from the 2005 corpus. This training data however did not include all examples

of all of the broadcast news channels found in the 2006 collection.

The primary evaluation metric for High-Level Feature Detection is Inferred Average Precision (InfAP). InfAP is similar to mean average precision (MAP) in that it measures both precision and recall whilst taking into account rank position, but varies in that it makes use of sampled truth data, rather than complete truth data. More information can be found in [32].

1.2. Terminology

In the High-Level Feature Detection tasks, several of the terms used can have multiple meanings. To clarify this we will use the following naming conventions for the remainder of this paper:

- *High-Level Feature Detection:* refers to the TRECVid task of classifying content against a set of semantic labels.

- *Low-level feature, or feature:* MPEG7 Visual descriptors, describing things such as colour or edges.

- *High-Level SVM:* Refers to the output of classifiers built on the outputs of our baseline SVM and of the K-Space Semantic Features, both defined in Section 3.

- *Semantic feature:* A high level description of content which goes beyond a description of colour or edges, but instead describes content in terms which infer some semantic description.

- *Concept:* The target of classification within the context of the High-Level Feature Detection task, will be one of the 39 concepts specified by NIST. Analogous to a topic in a search task.

2. RELATED WORK

Many systems that address the problem of High-Level feature detection treat the problem as a supervised pattern classification task [22]. Several overview papers which examine automatic concept detection have been written, notably works by Snoek [27], Naphade [16] [19] and Smeaton [22]. As stated in Smeaton [22], recent approaches to High-Level feature detection have utilized more generic frameworks as the task of building specific detectors for every type of semantic feature that might be required is not realistic.

Some of the earliest work to generically detect semantic features within digital video using statistical machine learning approaches was performed by Naphade and Smith [18]. Naphade and Smith used Support Vector Machines (SVM) and experimented with varying the number of annotations used for training as well as the machine learning kernel functions and associated parameters. This approach makes use of early fusion to combine multiple low-level features into a single representation. Using the TREC Video 2002 corpus, they achieved good performance in detecting some semantic concepts.

In earlier work, Naphade and Huang recognized that semantic features do not occur in isolation, and that semantic co-occurrence can be modeled to increase the probability of successful classification. For example, if a shot was successfully classified as containing ‘sky’ and ‘water’, the probability of a ‘beach’ being detected should be boosted and conversely the probability of the shot being ‘indoor’ should be reduced [15]. This approach which produced a generic framework for modeling semantic concepts was later extended by Naphade and Smith to incorporate their previous use of SVM’s for High-Level feature detection [17].

In 2003, a joint TRECVid submission from a combined IBM/Columbia University team established the pipeline analogy for feature processing by defining a ‘Concept Detection Framework’ which con-

sisted of multiple modules or 'silos', the sequencing of which was referred to as the 'Concept Detection pipeline' [5]. Four silos comprise the framework, the first to extract low-level multimedia features. The second silo creates unimodal feature representations, making use of multiple SVM's with different parametric settings, some of which are trained on multiple features aggregated through early fusion, whilst others are trained on single low-level representations. A validation set was used at this stage to select the best parametric settings for each feature being detected. The third silo uses multiple fusion approaches to combine the multiple models created in the second silo into a representation for each feature. Again validation sets were used for optimizing the fusion strategies. The final silo incorporates the earlier work by Naphade and Huang on semantic concept co-occurrence [15], as well as further statistical learning approaches, including SVMs, to create models which incorporate co-occurrence information. A final filtering phase is applied to improve precision, making use of collection knowledge such as down-weighting content from C-SPAN (a broadcast channel that only broadcasts political news or live sessions of government) if the feature being filtered is for sports. The use of validation data allows this approach to select the best possible candidates for representation of a feature from the many different models and permutations that are created, and as such achieved high performance. Furthermore, as stated earlier, this approach through its extensive use of machine learning techniques allows it to handle many different types of semantic features.

Snoek et. al. extend the pipeline approach, which utilizes semantic context through co-occurrence, to incorporate not only content, context data, but also style information from video [28]. Snoek et. al. observe that produced video (an example of which is the broadcast news used in TRECVid), is the end product of an authoring process which makes use of established conventions and techniques in order to emphasize parts of the content. Snoek defines the 'Semantic Pathfinder' a framework that contains

three conceptual modules for analysis of multimedia from a content, style and context perspective. Like previous approaches here, the Semantic Pathfinder makes extensive use of SVM's and emphasizes the use of validation sets for parameter tuning. The major innovation in this work is the style analysis step, which aims to analyze the roles of the editor, production design, production recording unit and preproduction team.

A common theme among the previous approaches to feature detection presented here was that the majority of the work was conducted within the framework of the TRECVid evaluations. As such TRECVid can be seen as a key driver of research advancement in this area [5], [28], [19], [12]. The recent conclusion of the 2006 TRECVid workshop demonstrated current best practice in the field of feature detection. Here we will briefly summarize the approaches of the top-performing groups at TRECVid 2006.

Tsinghua University's Intelligent Multimedia Group achieved the best performance for high level feature detection in TRECVid 2006 [7]. Their best run employed a weight and select fusion algorithm to select the top 50 classifiers to fuse from a set of 110 classifiers. They employed a hierarchial ensemble architecture that combined low-level features with diversified classifiers for each feature. Whilst highly effective this approach is very computationally expensive.

The MediaMill team achieved high performance in this task[26] utilizing their previously defined 'semantic pathfinder' [28]. MediaMill experiment with using variants on standard SVM approaches for supervised learning, making use of logistic regression and Fisher's linear discriminant. Whilst these later two generally performed worse than an SVM approach, they had the advantage of being computationally less expensive and required no parameter tuning. MediaMill emphasized the use of both early and late fusion strategies, however they found that whilst fusion generally provided improvements, what is to be fused needs to be carefully selected, as fusing all experiments for each concept being detected reduced

performance.

The Informedia group from Carnegie Mellon University reiterate the importance of cross-validation for parameter tuning of SVMs for each feature to be detected, noting the considerable difference in performance between optimized and default parameters [11]. Informedia use four low-level features and train a classifier for each of these visual features for each concept to be detected. Fusion of these classifiers was performed through logistic regression. Similar to other groups, Informedia exploited the conceptual links between semantic features through statistical machine learning and a probabilistic framework.

A common theme in more recent work presented here is the increase in the number of models being used for detection, with many approaches utilizing multiple models and performing cross-validation in order to select the best. Whilst highly effective, these approaches are also computationally expensive in terms of the sheer number of models which are created in order to select the best. Our work on the other hand which employs statistical machine learning makes use of only modest parameter searching and in some ways can be seen as being analogous to the early work by Naphade and Smith in SVM utilization for feature detection. However our distinguishing feature in this work is our approaches to late fusion of related concepts. One of these approaches is an application of the revised Dempster-Shafer theory, the transferable belief model. The second approach is work derived from our techniques for late fusion in the search domain which examines the distribution of scores for a given set of features. Based upon empirical observations of a correlation which appears to exist between the shape of the a feature's distribution and its relative performance compared against the other features in the set, this technique will produce dynamic weights which should favor the relatively better performing feature.

3. FEATURE DETECTION APPROACHES

We will define three approaches, a baseline SVM approach, and two alternate methods that add in new information to the initial baseline ranking. The additional information we will be adding will be specific feature detectors that were developed as part of the K-Space[1] team's participation in TRECVid 2006. This data to be added will either be as an individual feature, or as the output of a set of SVM's trained on the complete set of these features and of the outputs of the original baseline SVM models. The following text briefly highlights the set of K-Space Semantic Features and the High-Level SVMs. A detailed description can be found in the K-Space TRECVid 2006 notebook paper [30].

There were multiple semantic features developed by K-Space partners which formed key parts of the K-Space TRECVid submission. The Technical University of Berlin (TUB) supplied face statistics which detailed the size of the largest face present in a shot, as well as the number of faces within a shot. Outdoor detection was provided by Institut Eurcom and was based upon colour and texture low-level features extracted from regions, which are then classified through an ensemble of machine learning approaches and combined. The Image Video and Multimedia Laboratory, National Technical University of Athens provided 7 semantic features, desert, vegetation, mountain, road, sky, explosion and snow. This approach makes use of segmented regions to produce a 'Region Thesaurus'. Queen Mary University London (QMUL) also produced 4 semantic features for boat, US Flag, Weather and Maps. The QMUL approach utilized a machine learning ensemble which first employed a high-recall layer, followed by a high-precision layer.

[1]K-Space is an EU-funded Network of Excellence which brings together 14 partners from throughout Europe and which participated in TRECVid 2006 as a singe site

The High-Level SVM we produced was trained on the output of the aforemention classifier outputs, as well as the outputs from our baseline SVM classifiers. It was a basic second layer approach, and did not leverage any co-occurrence information or semantic associations. Improvements could be made to these set of classifiers by being more discriminatory in choosing what to combine for a given concept.

3.1. Baseline SVM

Our baseline run was generated by making use of SVM's which were trained on the common annotation set of the TRECVid 2005 corpus. The SVM implementation we used was svm_light [2]. For each of the 39 concepts to be classified we assigned an SVM to each classification task.

As input into the SVM's we used 6 different low-level visual descriptors. The low-level visual features we used are MPEG7 features and were extracted using the aceToolbox, developed as part of our collaboration in the aceMedia project [3]. Each NRKF keyframe that aligned with the annotation data was processed and included in the training set. These MPEG7 descriptors were Colour Layout, Colour Moments, Statistical Texture, Homogenous Texture, Edge Histogram and Scalable Colour. A complete description of each of these descriptors can be found in [20].

These features were normalized into range [-1:1] and aligned with the common annotation set on the training data from the 2005 corpus. Early fusion was then performed to create a single representation vector for each keyframe. Our SVM parameter optimization was ad-hoc, with a partial exploration of the parameter space. We explored the use of various kernels including linear and polynomial, finally settling on the radial basis function (RBF) as our kernel.

Our SVM's were run on a variety of hardware from multi-core servers, through to desktop machines. The slowest machine used was a Pentium 4 2.4 Ghz desktop. Training times for each of our SVM's did

not exceed 20 hours on our slowest machines.

Classification was performed against the 2006 test collection, specifically the NRKF keyframes of the test collection. In this case there were 75,000 NRKF keyframes that comprised the training collection, and 145,000 for the test collection, representing a much larger test than training collection.

At the end of this process, the NRKF results were aggregated back to the shot level, making use of a MAX function when assigning scores to shots.

3.2. Dempster-Shafer Combination

Of our six submissions to feature detection in TRECVid 2006, three were submissions which used the fusion of the outputs of other runs. For two of these runs we combined our baseline SVM data with several of the specialized semantic feature detectors provided by our K-Space partners using *Dempster-Shafer* (DS) theory [9, 21, 24, 23], which offers a convenient model for fusing imprecise and uncertain information.

3.3. Application of DS theory to the problem

Our approach to integration of evidence from multiple feature detectors using DS theory is primarily based on revised version of DS theory proposed by Smets [24, 23] called *Transferable Belief Models* (TBM). This section describes the major aspects of the application of the TBM theory to the problem of integration of evidence from multiple semantic feature detectors such as: (i) definition of the *frame of discernment*, (ii) general form of the belief structure used to model the way a piece of evidence is brought by a source to a proposition, (iii) taking into account source reliability, (iv) combination of beliefs from multiple sources, and (v) ranking of shots based on combined beliefs.

Let the *frame of discernment* Ω represent a set of two exclusive and exhaustive hypotheses, one that

a given concept is present and other that a given concept is absent in the keyframe: $\Omega = \{PRESENT, ABSENT\}$. The power set 2^{Ω} of Ω is composed of 4 propositions A: $2^{\Omega} = \{\ \emptyset, \{PRESENT\}, \{ABSENT\}, \{PRESENT \cup ABSENT\}\}$. The proposition $\{PRESENT \cup ABSENT\}$ will be also referred to as doubt ($DOUBT$).

Let us also define a set of U independent evidence sources (i.e. semantic feature detectors) S_1,..., S_U. A piece of evidence (measurement) C_u brought by a source S_u to a proposition A is modelled by the belief structure m_u called the *Basic Belief Assignment* (BBA) formally defined as: $m_u : 2^{\Omega} \longrightarrow [0, 1]$ with $m_u(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m_u(A) = 1$. In other words, for each measure C_u model m_u maps each value of C_u into a belief on a proposition (singleton or composed hypothesis) of 2^{Ω}.

In our approach the BBA models for each detector to be fused are based on both expert knowledge and statistical knowledge obtained from the development collection using methodology similar to the one proposed in [10] for training of facial emotion classifiers.

The general form of BBA models is shown in Figure 1a. Let us denote the neutral value of measure C_u as T_u^c, i.e. the value of C_u which does not provide evidence either for presence or absence of the given concept and for which the entire mass of belief is associated with doubt ($PRESENT \cup ABSENT$) while the masses of belief associated with propositions $PRESENT$ or $ABSENT$ are equal to zero. As the value of C_u decreases (increases) from T_u^c to T_u^l (T_u^r) the mass of belief associated with doubt decreases while the mass of belief associated with $ABSENT$ ($PRESENT$) increases. Finally, for values of C_u below (above) T_u^l (T_u^r) the entire mass of belief is associated with $ABSENT$ ($PRESENT$) and the mass of belief associated with doubt is equal to zero.

When a source is considered not completely reliable the confidence in this source can be attenuated [24]. Figure 1b shows an example of the BBA from Figure 1a discounted by factor $\alpha_u = 0.3$.

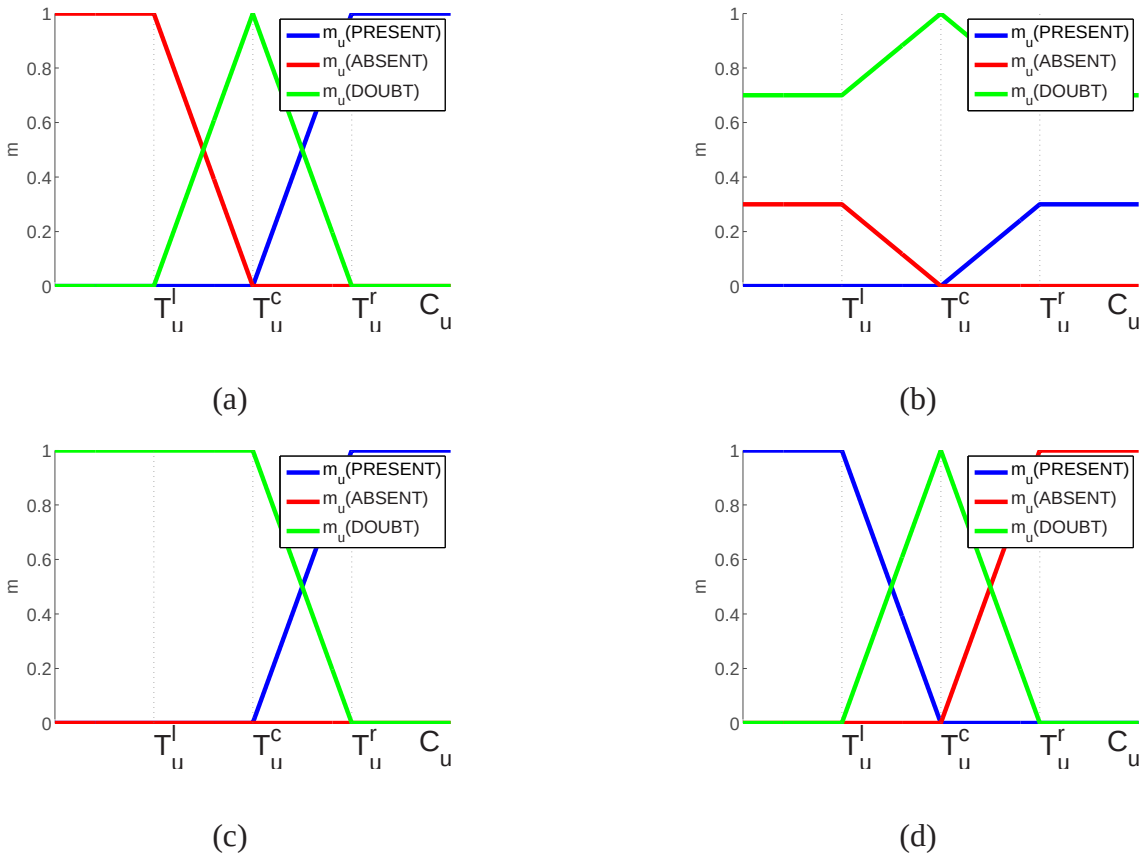

Fig. 1. General form of BBA functions.

It should be observed that discounting simply transfers the belief from propositions $PRESENT$ and $ABSENT$ to $PRESENT \cup ABSENT$ (doubt).

The above form of BBA can model in an intuitive way situations when fused sources provide evidence for equivalent concepts (e.g. fusion of two or more *"mountain"* detectors). In other words, this form of BBA is suitable for cases where a source is allowed to provide evidence for both hypotheses (i.e. some values of measurements provided by the source support the presence and others the absence of a given concept). However, very often there is a need to model situations where a source is allowed to provide evidence supporting only one of the two hypothesis (presence or absence of a given concept), e.g. presence of *"mountain"* provides strong evidence for *"outdoor"*, but absence of *"mountain"* does not necessarily indicate absence of *"outdoor"*. An example of a form of BBA which can be used to model

such cases is shown in Figure 1c[2]. Also, situations where presence(absence) of one concept indicates absence(presence) of another concept (e.g. a trivial case where presence of *"indoor"* provides evidence for absence of *"outdoor"*) can be easily implemented using the form of BBA from Figure 1d. Of course many of the above cases may occur together. For example the presence of one concept may indicate the absence of another concept but absence of the first concept does not necessarily indicate presence of the other (e.g. *"snow"* and *"desert"*). In such cases, the required forms of BBA can be constructed by combining appropriate transformations of the BBA from Figure 1a.

All the parameters required for each BBA, including the form of BBA used, thresholds T_u^l, T_u^c, and T_u^r and discounting factor α_u are discovered through experimentation on the training collection as described in the next section.

Once the BBA for each source is defined the evidences from multiple sources can be fused using a commonly used operator called the orthogonal sum or Dempster's rule of combination [9, 23, 24]. According to this operator, which is commutative and associative, the combined belief structure $m^{\oplus}$ is defined by:

$$m^{\oplus} = m_1 \oplus m_2 \oplus \ldots \oplus m_U \tag{1}$$

where for two sources of information S_u and S_v the combined belief structure $m^{\oplus}$ is defined as:

$$\forall A \subseteq \Omega \qquad m^{\oplus}(A) = \sum_{B \cap C = A} m_u(B) \cdot m_v(C) \tag{2}$$

Finally, all shots are ranked according to the combined beliefs that a given concept is present or not in the shots using an empirically derived formula:

$$C_{Total} = m^{\oplus}(PRESENT) - m^{\oplus}(ABSENT) \tag{3}$$

[2]Note that this BBA is equivalent to BBA from figure (a) after discounting all values $C_u < T_u^c$ by factor $\alpha = 0$

where $m^{\oplus}(PRESENT)$ and $m^{\oplus}(ABSENT)$ are combined beliefs that measurements obtained for the given shot from all integrated sources of evidence support the hypothesis $PRESENT$ and $ABSENT$ respectively.

3.4. Designing Belief Structures for each Source of Information - Training

For each semantic feature the design of BBAs for each source of information S_u to be fused involves selection of the form of BBA, selection of the thresholds T_u^l, T_u^c, T_u^r and discounting factor α_u, all of which are estimated using a training collection of manually annotated keyframes.

For each source to be fused S_u the thresholds T_u^l and T_u^r were estimated by finding respectively the minimum and maximum values of C_u in the training population of shots. The value of T_u^c was computed simply as the midpoint of T_u^l and T_u^r. For simplicity, the discounting factor corresponding to the most reliable detector was set to one. For the remaining sources of evidence discounting factors were found by combining them with the most reliable detector and searching for discounting factor/factors minimizing the MAP. Also the form of each BBA was selected by searching for the form maximizing the value of MAP for a given concept to be detected.

3.5. Score Distribution Analysis

The second approach to fusion for building classifiers that we used is based on a dynamic weighting function, which is designed to infer relative performance given a set of input features. For example, we use the fact that feature *A* will perform better than feature *B*, as opposed to predicting that feature *A* will achieve *X* performance and weight accordingly, i.e. it is not design to infer absolute but *relative* performance of a feature.

The central thesis of the approach here is that by examining the distributions of the scores generated

from a feature, it is possible to infer relative performance of one feature against another. This principle was developed previously within the context of multimedia search, where the combination of various pieces of evidence is crucial in attaining decent levels of performance. Details of our work in the search domain utilizing this method can be found in [31]. Before proceeding with an explanation of this approach, we will first define what we are attempting to fuse together in order to enhance performance of High-Level Feature Detection.

Two approaches to feature detection will be used for fusion in this approach. The first of these is our baseline SVM, as detailed in Section 3.1. The second approach we use is what we refer to as a 'High-Level SVM'. This is an SVM which is trained on the outputs of the baseline SVM along with the outputs of a series of specific feature detectors developed by our K-Space partners as part of our TRECVid collaboration. Specifically, this SVM made use of the following custom feature detectors: Boat, Building, Crowd, US Flag, Map, Weather, Desert, Fire, Mountain, Road, Sky, Snow, Vegetation and Face. When combined with the outputs of the 39 low-level SVMs, this produced a feature vector of 53 elements. More information on these detectors can be found in the K-Space TRECVid notebook paper [30]. Therefore it was the fusion of these two sources that our score-based approach was applied to.

We have previously observed a correlation in the search domain where if a feature undergoes a rapid change in its normalized scores, then that feature is likely to perform better than a feature which undergoes a more gradual transition in normalized scores. Figure 2 illustrates this correlation.

The results for this feature with respect to InfAP are given in Table 1. The baseline SVM, which is the feature which undergoes the greater initial change in normalized score, is also the feature which achieves the best performance.

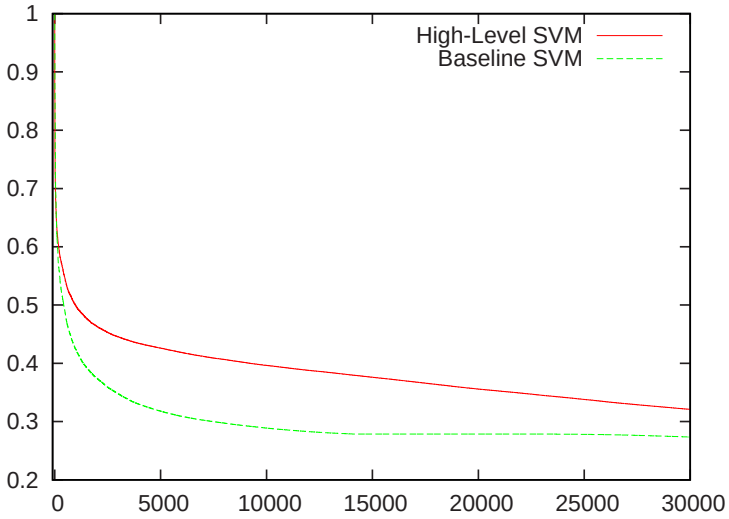

Fig. 2. Waterscape

	Baseline SVM	HighLevel SVM
InfAP	**0.1316**	0.0806

Tab. 1. Waterscape Feature Resutls

As a second example we can present this same correlation with the output of the classification for ‘People Marching’ as demonstrated in Figure 3.

The results for ‘People Marching’ are given in Table 2. Again, the better performing feature in this case is the Baseline SVM, and it’s curve of results is steeper than the HighLevel SVM.

	Baseline SVM	HighLevel SVM
InfAP	**0.0282**	0.0222

Tab. 2. People Marching Feature Results

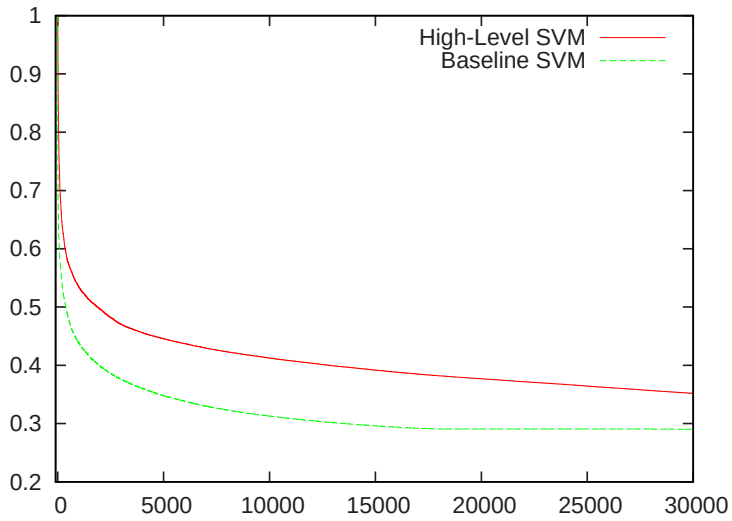

Fig. 3. PeopleMarching

We have hypothesized that the reason this correlation exists within the search domain is that the rapid change in score can be seen as an indicator of 'interesting-ness' or having found multiple artifacts, that whilst all different are important enough to have been promoted to near the top of the ranking. Conversely, a feature which exhibits a more gradual change could be thought of as having found many artifacts that are very similar, and as such has not been able to differentiate these to a large degree from the set. However a caveat to this thinking is that these correlations are more likely to exist with 'high-noise' data, and that well performing features may not exhibit the same correlations as they will have likely found greater quantities of similar highly ranked elements.

This line of thinking is derived from observations made by Lee [14] where he states that fusion appears to work because "different runs might retrieve similar sets of relevant documents but retrieve different sets of non-relevant documents". Our work in applying these observations to the task of High-Level Feature Detection opens up many possible avenues for exploration, and whilst there are many overlaps in the task of search versus feature detection, such as creating a ranked list of results, there are a

few differences as well. We have also noted that the above correlations are not universal in this collection, and there are instances in which the correlation does not hold. Figure 4 highlights a failure case for our above hypothesis, with results given in Table 3 which demonstrates that whilst the better performing feature was the baseline SVM, it did not exhibit the same shape in curve as previous examples.

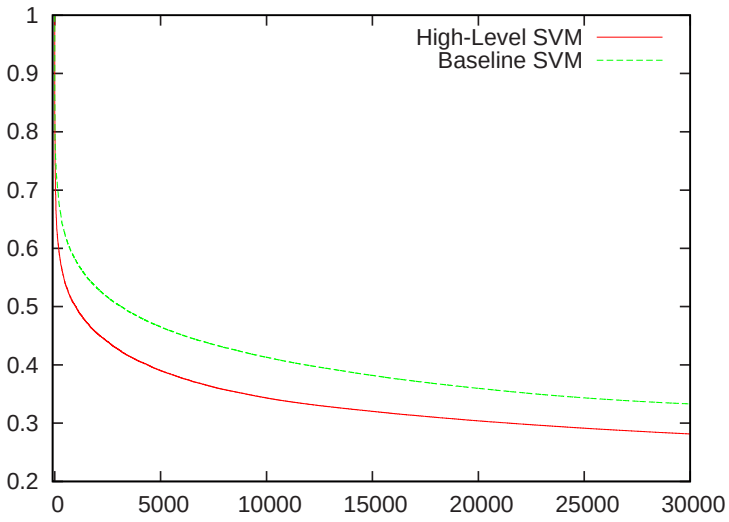

Fig. 4. Meeting

	Baseline SVM	HighLevel SVM
InfAP	**0.1788**	0.1171

Tab. 3. Meeting Feature Results

Nevertheless, if we proceed with the assumption that this correlation holds for some of the features, we can attempt to exploit this property to dynamically generate weights to fuse these sources of information together. In order to combine these sources of evidence, we first normalize the scores using MinMax normalization, as given by in (4).

$$Norm_{score(x)} = \frac{Score_x - Score_{min}}{Score_{max} - Score_{min}} \tag{4}$$

Next we calculate the average change in score for a given set size of a feature, which we refer to as the Mean Average Distance (MAD) (5).

$$MAD = \frac{\sum_{n=1}^{N}(score(n) - score(n+1))}{N-1} \tag{5}$$

A direct comparison of this average score distance between features would not necessarily work as it may not account for differences in scoring metrics that are used, or by the natural distribution of a feature amongst its scores. In order to make a cross-comparable feature value we compute a ratio of MAD of a top subset of that feature, versus a larger set of that feature. In this series of experiments we examined the top subset size of 5% of the feature over 95% of the feature. The value this produces we refer to as a Similarity Cluster (SC) value. This is formally defined in (6).

$$SC = \frac{MAD(subset)}{MAD(largerset)} \tag{6}$$

Once we have this value for a given feature, we can then use it to generate a weight for that feature, as shown by 7.

$$Feature\ Weight = \frac{Feature\ SC\ Score}{\Sigma All\ SC\ Scores} \tag{7}$$

This formula will generate larger weights for features which exhibit larger values of SC, in turn meaning that these features underwent the greater initial score change and according to our observations these features are likely to be the better performing features and should be weighted accordingly.

3.6. Motivation for fusion strategies

The selection of our two fusion strategies for experimentation was to explore two different themes. The first, addressed by the Dempster-Shafer fusion approach, is that if we know something about the information to be combined, and something about its quality, that we should be able to exploit this through the Dempster-Shafer framework which allows this additional data to be explicitly modeled. The second theme, addressed by the score based approach, is that we may not always have reliable information about the quality of an information source, and as such are there ways in which we can combine the information which can dynamically infer if a source is likely to be perform better or worse relative to what it is being combined with.

We chose to investigate the application of Dempster-Shafer as it allows experts to easily use their knowledge about the complex relations between fused concepts, and represents a powerful tool for combining measures of evidence. The main features of Dempster-Shafers theory include: the convenient management of uncertainties by explicitly modeling the doubt, the ability of taking into account reliability of information sources, distinct representation of ignorance, imprecision and conflicting information and convenient incorporation of statistical and/or expert knowledge. In other words, it appears more general and more flexible than the commonly used probability model and allows representation of states of knowledge and opinions that probability models cannot [10].

Motivation for the application of the score based fusion approach stemmed from the need to dynamically generate classifier coefficients to allow for weighted classifier combination. Whilst this data could be derived from training examples it would be reliant on the training examples being fully representative of the test data. Furthermore this investigation of score based combination is not an exploration of determining the absolute performance of a classifier. Rather it is concerned with determining rela-

tive performance of the classifiers to be combined so that appropriate coefficients can be determined to maximize the chances of successful fusion.

We believe that for both of these approaches that they will offer advantages over traditional approaches to evidence combination such as summing or multiplying the classifier outputs. In the case of Dempster-Shafer fusion, it allows for evidence merging which cannot be sufficiently modeled through summing or multiplication. For example, an information source can provide evidence supporting only one of two hypothesis (presence or absence of a concept), such as the presence of 'mountain' provides strong evidence for 'outdoor', but the absence of 'mountain' does not necessarily indicate the absence of 'outdoor'. Score based fusion allows for the dynamic generation of coefficients for fusion. Methods which employ summing or multiplication must have some prior concept of performance to employ some weighting scheme as often equal weights for fusion are not desirable.

From a computational perspective we believe these two fusion approaches to be faster to calculate than those of SVM-based combination. Whilst our fusion approaches do not perform as well as SVM-based combination, our approaches are faster to compute as we are performing various transformations of the existing classification scores through the application of coefficients, as opposed to generating new classifications. However as we will subsequently note, we would not see the two approaches as being mutually exclusive.

4. ANALYSIS

In this section we will cross-compare the results of the three runs with respect to specific features. The first half of these experiments were first presented in [29]. This section will perform an in-depth analysis of the results these three runs of the High-Level Feature Detection task from TRECVid 2006. The first

of the three runs is our Baseline SVM submission, trained using low-level MPEG7 visual features. The second run is our Dempster-Shafer (DS) fusion run, which fused the output of the Baseline SVM with a specific feature detector provided by one of our K-Space partners. The final run is our score distribution fusion method, which for a given concept, combines the output of the Baseline SVM with the output of the High-Level SVM, that is an SVM that has been trained on the outputs of the Baseline SVM and with the specific feature detectors. It should be noted that for each concept being detected one of the inputs was always the Baseline SVM. We will present in this section a detailed evaluation of these approaches by examining the output of results of several specific target features, and compare the results of the Baseline SVM against one of the fusion methods.

To do this we will be using five different evaluation metrics. The first of these is InfAP which was the official TRECVid metric used for ranking. The second metric is Precision at 100 (P@100), which measured the amount of relevant shots seen after an examination of the top 100 results for a given list. The remaining three metrics examine how the various result sets overlap, to provide indications if the sets are returning similar or different rankings. Our first basic overlap measure is an intersect statistic, which computes the degree to which two sets overlap, given as:

$$\frac{Run\_A \cap Run\_B}{Run\_A \cup Run\_B} \tag{8}$$

The second and third measures will we use are called R-Overlap and N-Overlap, representing the relevant set overlap between results sets, and the non-relevant set overlap between results sets. First given in Lee [14] and refined later in [6], these formally are:

$$R_{Overlap} = \frac{R \cap S_1 \cap S_2 .... \cap S_n}{(R \cap S_1) \cup (R \cap S_2) \cup ...(R \cap S_n)} \tag{9}$$

$$NR_{Overlap} = \frac{NR \cap S_1 \cap S_2 .... \cap S_n}{(NR \cap S_1) \cup (NR \cap S_2) \cup ...(NR \cap S_n)} \quad (10)$$

where *R* means relevant, *NR* non-relevant and *S* are the result sets being compared.

Table 4 presents the results for each run averaged across all topics. Whilst not ideal, this averaging produces figures which represent the relative effectiveness of each of our approaches. The Baseline SVM remains the best performing feature, followed by our Score Based approach and finally the DS approach.

	Baseline SVM	DS Fusion	Score Based Fusion
InfAP	**0.1104**	0.0849	0.0977

Tab. 4. Overall Results

This averaged InfAP measure, whilst demonstrating that overall the Baseline SVM approach performs the best, obscures details contained within individual detected features, where each approach had varying degrees of success. A complete overview of the results of each fusion method for every feature evaluated can be found in [30].

In Table 5 we examine eight detected concepts, four each for both of our fusion methods, and compare the results against the results for the baseline SVM. The Table can be read as follows, the first column defines the feature being detected. The second column is the InfAP score for the Baseline SVM. The third column presents the second information source that is being used for fusion (either from the HighLevel SVM or a specific feature detector). In the fourth column we specify which of the two fusion approaches was used. The fifth column is the InfAP for that fusion result. The sixth, seventh and eighth columns are the values for R-overlap, NR-overlap and intersection of the fusion output compared against

INPUT			FUSION					
Concept	Baseline	Feature	Type	InfAP	R-overlap	N-overlap	Intersection	Unique
PeopleMarching	0.0282	HighLevel	Score	0.0381	0.90	0.72	0.72	8
Maps	0.2484	HighLevel	Score	0.2437	0.92	0.54	0.57	14
Airplane	0.0105	HighLevel	Score	0.0201	0.66	0.39	0.40	18
PoliceSecurity	0.0146	HighLevel	Score	0.0154	0.80	0.62	0.62	9
Desert	0.0588	Desert	DS	0.057	0.88	0.73	0.74	3
Mountain	0.0546	Mountain	DS	0.0548	0.73	0.61	0.62	16
ExplosionFire	0.0679	Face	DS	0.0734	0.91	0.84	0.85	9
Military	0.0733	Face	DS	0.0636	0.73	0.61	0.62	27

Tab. 5. Specific Feature Detection Results

the baseline SVM output. The final column is a count of the number of unique relevant shots that the fusion model found that were not included in the baseline SVM results.

A note to make on these figures is that in the majority of cases, the results being fused with the baseline often performed worse than the baseline, in some cases this performance degradation exceeded 100% (for example the High-Level SVM score for Maps was InfAP 0.1196).

The hardest data to interpret from Table 5 is the impact of the R-Overlap and NR-Overlap scores. The reasoning for including this data is to begin to examine what constitutes good candidates for fusion. To simplify the task of processing these values, we have plotted these in Figure 5. The ‘X’ axis represents the relevant overlap and the ‘Y’ axis the non-relevant overlap. The key in the top left includes in brackets for each of the previous results, what each fused concept scored in terms of InfAP and how much change there was from the baseline (i.e. positive meaning the fused results improved performance). We have

included a line on the graph which represents an even split between relevant and non-relevant overlap, such that any point on that line would be said to have an equal amount of overlap of relevant shots and non-relevant shots. On Figure 5 all points are to the right of the line, indicating that there is a greater proportion of relevant overlap than non-relevant overlap. However, we need to be careful when interpreting this graph as the axis on the graph are proportional. The absolute number of relevant shots is far less than the non-relevant. Repeated experiments will be required to draw any firm conclusions from the overlap data.

We can hypothesize about potential beneficial overlap characteristics from this graph. For instance, a high relevant overlap between two results sets to be fused would indicate that not much new information will be included by fusing those result sets. However if those same two result sets have a low non-relevant overlap, then the relevant shots should be promoted higher in the fused result set because of the overlap. Alternatively a lower relevant overlap would indicate that more new relevant information would be included by combining these two result sets. In either case, minimizing the non-relevant overlap would help to promote relevant shots higher up in a fused result set.

There are many observations to be made from these results. The first note to make is that for several of the concepts our fusion methods achieve a higher InfAP than that of the baseline, whilst at time returning different relevant shots. The second observation is that in the many of the fusion scenarios, there was a large discrepancy between the quality of the baseline SVM results and what it was being fused with. For instance, for the ‘Maps’ feature, the baseline SVM performed well with an InfAP of 0.2482, whilst the ‘High-Level SVM’ result obtained an InfAP of 0.1196. Despite this the fusion method utilized produced a result of 0.2437, comparable with the baseline result. Furthermore, the fusion method introduced 14 different relevant shots than what the baseline SVM had found, whilst returning a significantly different

result set, with only 57% of shots shared between the two.

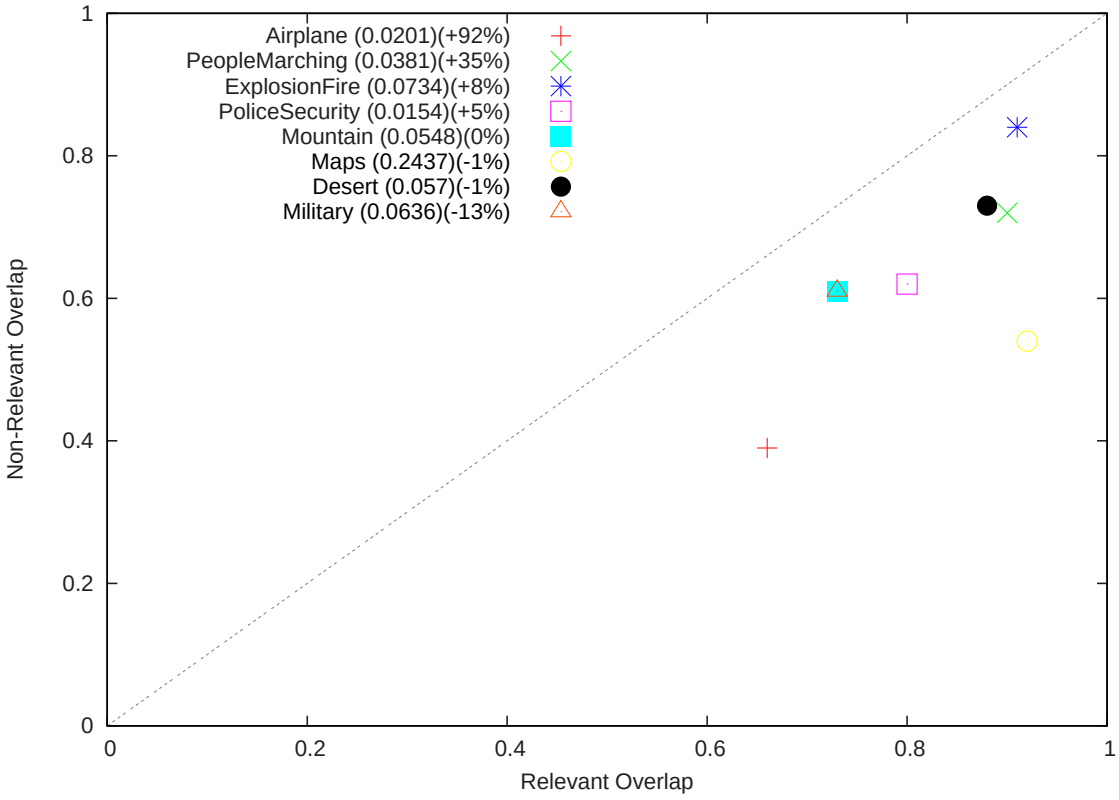

Fig. 5. Overlap Visualization

Of the results presented here, the fusion methods achieve comparable results to the baseline SVM, returning similar sets of relevant shots. However included in the set of relevant shots are shots which were not detected by the baseline SVM, increasing the diversity of shots found. Closer inspection of the results reinforces Lee's hypothesis that the fusion will result in the bringing together of similar sets of relevant documents and different sets of non-relevant documents. This can be observed through the differences in values of R-overlap versus N-overlap, where N-overlap generally is a much smaller value than R-overlap.

The results shown in Table 5 were features that were included in the judgement pool by NIST for evaluation in TRECVid. Out of the 39 features to be detected NIST evaluated 20, however participation

required that all 39 features be attempted. Further to these initial experiments which we presented in [29], we present new results for which we performed our own relevance judgements to further examine our approaches. These results are presented in Table 6, with results given in terms of Precision at 100 (P@100). For reference we also include the results from the High-Level SVM, and as noted previously the score based fusion method was a fusion of the baseline and High-Level SVM results.

Concept	Baseline	DS Fusion	Score Based	HighSVM
Court	0.24	0.24	0.27	0.01
Vegetation	0.80	0.81	0.82	0.83
Sky	0.90	0.93	0.90	0.89
Snow	0.29	0.29	0.27	0.09
Urban	0.41	0.38	0.41	0.35
Crowd	0.74	0.42	0.75	0.72
Face	0.97	0.99	1.00	0.87
Bus	0.02	0.02	0.02	0.01
Boat	0.08	0.08	0.08	0.02
Walking Running	0.18	0.06	0.14	0.18

Tab. 6. P@100 results for non-NIST judged features

The results in Table 6 generally achieve greater performance than those which were evaluated by NIST, most likely as they can be seen as ‘easier’ concepts. Within this context we can see smaller variations in performance than those presented previously. This could be due to the fact that for the most part, these features were far more accurate than others we evaluated, and as such may have less influence from noise in the results. One remarkable result we encountered was the score based fusion

approach for ‘court’. The ‘court’ baseline (performance P@100 of 0.24) was combined with the high-level SVM, which for this feature obtained a P@100 value of only 0.01. Despite this the approach was able to successfully fuse together the two features and attain an information gain that was greater than the sum of just the high-level SVM performance by itself. Whilst this result appears impressive, we need to perform further careful analysis to understand exactly what caused this result. This will refine our thinking as to what the a priori conditions need to be for successful fusion using this score based approach.

We can further observe here that for the score distribution fusion method, the largest increases in performance are for those features in which high levels of noise are present, which would appear to be in keeping with our earlier stated hypothesis about the likeliness of success for this approach. Further work we need to undertake is to attempt to establish some more concrete definition of what actually constitutes a high-quality information source versus a high-noise information source, as this would allow us to further develop our initial models, which at the moment make very basic assumptions about the quality of the representation. This would allow us to incorporate some form of threshold where if the quality of the sources were high, we could default to a different weight generation strategy, and vice versa.

We should also note that there are also cases where the fusion methods failed to achieve close to the baseline result, and merely introduced significant noise into the result set. These concepts included for the score based approach, ‘explosion (InfAP 0.0029)’, ‘charts (InfAP 0.0403)’ and ‘truck (InfAP 0.0253)’ concepts, whilst for the Dempster-Shafer fusion approach, features ‘meeting (InfAP 0.0277)’, ‘computer_tv_screen (InfAP 0.0237)’ and ‘people marching (InfAP 0.0026)’ proved particularly problematic [30]. We hypothesize that the Dempster-Shafer fusion, was over-reliant on the training data, and

thus we over fitted our weights for several of these concepts. This observation was reinforced by some of the failure cases we encountered in our second set of experiments presented in Table 6. Further work would be to employ more robust training schemes to avoid these over-fitting problems, and re-evaluate the effectiveness of this combination approach.

In terms of improvement left to be gained from these fusion approaches, we believe that considerable amounts of improvement can still be gained theoretically, despite the quality of several of the outputs being fused. This statement is derived from the observation that for all the outputs presented, none had an ideal ranking where no non-relevant shots appeared at the beginning of the result lists. As we are using InfAP, which favors results which place relevant before non-relevant shots, as one of our primary metrics, performance gains can be expected if we can produce results which eliminate non-relevant shots from the start of the result sets.

Each of our SVM approaches we believe also suffered from over training. As the development data for the 2006 was not fully representative of the 2006 test data, the risk of over training was significant. Given the experience of other groups in the task and a re-examination of our training methodology, we believe that both of our SVM approaches suffered from over fitting on the development data. This produced a knock-on effect on our training of the fusion approaches, which implicitly would have incorporated the over fitting. The positive to be seen from this is that we can address these issues to obtain performance gains for each fusion approach through regenerating our SVM classifiers.

The fundamental observation that we can draw from this set of experiments is that the fusion approaches we have identified here were for several features, able to achieve performance comparable or better than the baseline SVM whilst returning a different set of relevant shots. This means that despite the fusion of a relatively high quality information source (baseline SVM) with a low quality information

source (high-level SVM) new relevant information was able to be found. It also means that there remains a significant space in which to explore such that we can aim to return the superset of relevant shots found from each information source.

5. CONCLUSIONS

In this paper we have presented attempts at extending the use of machine learning approaches to High-Level Feature Detection through the application of two fusion algorithms. These algorithms are designed to be inexpensive extensions to machine learning approaches and can be seen to add novel results to existing result sets without the need for computationally expensive training. Furthermore we have shown that in some cases novel results can be found through the fusion of a high-quality information source with a lower quality source, at minimal expense to overall performance. We note that we do not view these approaches as a replacement for existing High-Level Feature Detection techniques, but as complimentary approaches that can be used to increase the diversity of the results generated.

We have advocated the use of the R-overlap and NR-overlap [14][6] as a mechanism by which we can compare result sets to be fused. Whilst not a perfect measure it allows us to hypothesize about attributes are beneficial to successful fusion.

Our fusion methods presented in this paper did not achieve consistent results. Both of these approaches however were often combining information sources that were not of approximate quality, and as such experimentation where this was the case would be beneficial. Furthermore a natural extension of this work would be to use the score based derived coefficients as inputs into a Dempster-Shafer framework. Before we perform that though we need to isolate the effect each of these fusion approaches is having on performance.

6. ACKNOWLEDGMENTS

We would like to extend our thanks to our TRECVid K-Space partners who provided the specific feature detectors used throughout this series of experiments.

We are grateful to the AceMedia project (FP6-001765) which provided us with output from the AceToolbox image analysis tooklit.

The research leading to this paper was supported by the European Commission under contract FP6-027026 (K-Space) and by Science Foundation Ireland under grant 03/IN.3/I361.

7. REFERENCES

[1] TRECVID homepage, http://trecvid.nist.gov/.

[2] svm_light, available from http://svmlight.joachims.org/.

[3] The AceMedia Project, available at http://www.acemedia.org.

[4] Arnon Amir, Janne Argillander, Murray Campbell, Alexander Haubold, Shahram Ebadollahi, Feng Kang, Milind Naphade, Apostol Natsev, John R. Smith, Jelena Tesic, and Timo Volkmer. IBM Research TRECVID-2005 video retrieval system. In *TREC Video Retrieval Evaluation Proceedings*, 2006.

[5] Arnon Amir, Marco Berg, Shih-Fu Chang, Giridharan Iyengar, Ching-Yung Lin, Apostol Natsev, Chalapathy Neti, Harriet Nock, Milind Naphade, Winston Hsu, John R. Smith, Belle Tseng, Yi Wu, and Dongqing Zhang. Ibm research trecvid-2003 video retrieval system. In *TRECVID 2003 Workshop*, Gaithersburg, MD, November 2003.

[6] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder, and Nazli Goharian. Fusion of effective retrieval strategies in the same information retrieval system. *J. Am. Soc. Inf. Sci. Technol.*, 55(10):859–868, 2004.

[7] Jie Cao, Yanxiang Lan, Jianmin Li, Qiang Li, Xirong Li, Fuzong Lin, Xiaobing Liu, Linjie Luo, Wanli Peng, Dong Wang, Huiyi Wang, Zhikun Wang, Zhen Xiang, Jinhui Yuan, Wuije Zheng, Bo Zhang, Jun Zhang, Leigang Zhang, and Xiao Zhang. Intelligent Multimedia Group of Tsinghua University at TRECVID 2006. In *TRECVid 2006 – Text REtrieval Conference, TRECVID Workshop, Gaithersburg, Md., 13-14 November 2006*, 2006.

[8] Michael G. Christel and Alexander G. Hauptmann. The use and utility of high-level semantic features in video retrieval. In *CIVR 2005 - International Conference on Image and Video Retrieval, W-K Leow et al. (Eds.), LNCS 3568, pp61-70*.

[9] A. Dempster. Upper and lower probabilities induced by multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[10] Z. Hammal, A. Caplier, and M. Rombaut. A fusion process based on belief theory for classification of facial basic emotions. In *Proc. 8th Int'l Conf. on Information Fusion, Philadelphia, USA*, July 2005.

[11] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H Lin, R. Yan, and J. Yang. Multi-Lingual Broadcast News Retrieval. In *TRECVid 2006 – Text REtrieval Conference, TRECVID Workshop, Gaithersburg, Md., 13-14 November 2006*, 2006.

[12] Alexander G. Hauptmann and Michael G. Christel. Successful approaches in the trec video retrieval

evaluations. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 668–675, New York, NY, USA, 2004. ACM Press.

[13] Wessel Kraaij, Alan F. Smeaton, and Paul Over. Trecvid 2006 - an introduction. In *Proceedings of TRECVID 2006*, November 2006.

[14] Joon Ho Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM Press.

[15] Milind R. Naphade and Thomas S. Huang. A probabilistic framework for semantic video indexing, filtering and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.

[16] Milind R. Naphade and Thomas S. Huang. Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. *Neural Networks, IEEE Transactions on*, 13(4):793–810, 2002.

[17] Milind R. Naphade and John R. Smith. A hybrid framework for detecting the semantics of concepts and context. In *CIVR 2003 - International Conference on Image and Video Retrieval*, pages 196–205, 2003.

[18] Milind R. Naphade and John R. Smith. Learning visual models of semantic concepts. In *ICIP (2)*, pages 531–534, 2003.

[19] Milind R. Naphade and John R. Smith. On the detection of semantic concepts at trecvid. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667, New York, NY, USA, 2004. ACM Press.

[20] Noel O'Connor, Edward Cooke, Herve le Borgne, Michael Blighe, and Tomasz Adamek. The Ace-Toolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

[21] G. Shafer. A mathematical theory of evidence. *Princetown Univ. Press, Princetown New Jersey*, 1976.

[22] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[23] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2):191–234, 1994.

[24] P. Smets, E.H. Mamdami, D. Dubois, and H. Prade. *Non-Standard Logics for Automated Reasoning*. ISBN 0126495203. Academic Press, Harcourt Brace Jovanovich Publisher, 1988.

[25] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000.

[26] Cees G.M. Snoek, Jan C. van Gemert, Theo Gevers, Bouke Huurnink, Dennis C. Koelma, Michiel van Liempt, Ork de Rooij, Koen E.A. van de Sande, Frank J. Seinstra, Arnold W.M. Smeulders, Andrew H.C. Thean, Cor J. Veenman, and Marcel Worring. The MediaMill TRECVID 2006 semantic video search engine. In *Proceedings of the 4th TRECVID Workshop*, Gaithersburg, USA, November 2006.

[27] Cees G.M. Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[28] Cees G.M. Snoek, Marcel Worring, Jan-Mark Geusebroek, Dennis C. Koelma, Frank J. Seinstra, and Arnold W.M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1678–1689, 2006.

[29] Peter Wilkins, Tomasz Adamek, Alan F. Smeaton, and Noel O'Connor. Inexpensive fusion methods for enhancing feature detection. In *CBMI 2007 5th International Workshop on ContentBased Multimedia Indexing*, 2007.

[30] Peter Wilkins and et al. KSpace at TRECVid 2006. In *TRECVid 2006 – Text REtrieval Conference, TRECVID Workshop, Gaithersburg, Md., 13-14 November 2006*, 2006.

[31] Peter Wilkins, Paul Ferguson, and Alan F. Smeaton. Using score distributions for querytime fusion in multimedia retrieval. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[32] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111, New York, NY, USA, 2006. ACM Press.