Early versus Late Fusion in Semantic Video Analysis

Cees G.M. Snoek
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098 SJ
Amsterdam, The Netherlands
cgmsnoek@science.uva.nl

Marcel Worring
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098 SJ
Amsterdam, The Netherlands
worring@science.uva.nl

Arnold W.M. Smeulders
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098 SJ
Amsterdam, The Netherlands
smeulders@science.uva.nl

ABSTRACT

Semantic analysis of multimodal video aims to index segments of interest at a conceptual level. In reaching this goal, it requires an analysis of several information streams. At some point in the analysis these streams need to be fused. In this paper, we consider two classes of fusion schemes, namely early fusion and late fusion. The former fuses modalities in feature space, the latter fuses modalities in semantic space. We show by experiment on 184 hours of broadcast video data and for 20 semantic concepts, that late fusion tends to give slightly better performance for most concepts. However, for those concepts where early fusion performs better the difference is more significant.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms, Performance

Keywords

Multimedia understanding, early fusion, late fusion, semantic concept detection

1. INTRODUCTION

The promise of instantaneous semantic access to multimodal video repositories has triggered much attention for methods that automatically index segments of interest. Typical semantic video analysis methods first extract a set of features from the raw data. Choices here include a feature extraction method based on unimodal analysis, i.e. using only one information stream, or multimodal analysis, i.e. using two or more information streams. Based on these extracted features, an algorithm indexes the data with semantic concepts like *car*, *ice hockey*, and *beach*. At present, there is enough experimental evidence to state that semantic video analysis yields the most effective index when a multimodal approach is adhered [1, 3, 6, 9].

In general, three modalities exist in video, namely the auditory, the textual, and the visual modality. A multimodal analysis method for semantic indexing of video inevitably includes a fusion step to combine the results of several single media analysis procedures. Pioneering approaches for multimodal fusion focussed on indexing of specific concepts only, e.g. [7]. In these cases a rule-based combination method yields adequate results. Drawbacks of such approaches, however, are the lack of scalability and robustness. To cope with both issues, a recent trend in semantic video analysis are generic indexing approaches using machine learning [1, 3, 6, 9, 10]. As speech is often the most informative part of the auditory source, these approaches typically fuse textual features obtained from transcribed speech with visual features. We identify two general fusion strategies within the machine learning trend to semantic video analysis, namely: early fusion [6] and late fusion [1, 3, 9, 10]. The question arises whether early fusion or late fusion is the preferred method for semantic video analysis. In this paper, we discuss both multimodal fusion approaches and perform a comparative evaluation.

The organization of this paper is as follows. First, we introduce two general schemes for early and late fusion in section 2. Then we present an implementation in section 3. We discuss the experimental setup in which we evaluate both schemes in section 4. We present results in section 5.

2. FUSION SCHEMES

We perceive of semantic indexing in video as a pattern recognition problem. Given pattern x, part of a shot i, the aim is to obtain a measure, which indicates whether semantic concept ω is present in shot i. To obtain a pattern representation from multimodal video we rely on feature extraction. Early fusion and late fusion differ in the way they integrate the results from feature extraction on the various modalities. In the following description of the early fusion and late fusion scheme we assume that a lexicon of semantic concepts together with labeled examples exists.

2.1 Early Fusion

Indexing approaches that rely on early fusion first extract unimodal features. After analysis of the various unimodal streams, the extracted features are combined into a single representation. In [6] for example, we used concatenation of unimodal feature vectors to obtain a fused multimedia

MM'05, November 6–11, 2005, Singapore.

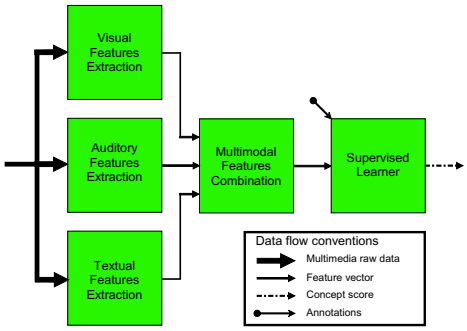

Figure 1: General scheme for early fusion. Output of unimodal analysis is fused before a concept is learned.

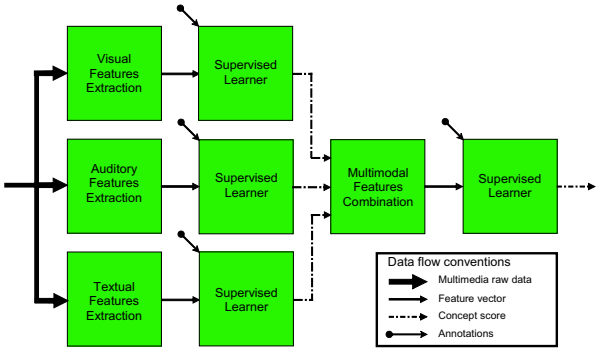

Figure 2: General scheme for late fusion. Output of unimodal analysis is used to learn separate scores for a concept. After fusion a final score is learned for the concept.

representation. After combination of unimodal features in a multimodal representation, early fusion methods rely on supervised learning to classify semantic concepts. We define:

DEFINITION 1 (EARLY FUSION). *Fusion scheme that integrates unimodal features before learning concepts.*

Early fusion yields a truly multimedia feature representation, since the features are integrated from the start. An added advantage is the requirement of one learning phase only. Disadvantage of the approach is the difficulty to combine features into a common representation. The general scheme for early fusion is illustrated in Figure 1.

2.2 Late Fusion

Indexing approaches that rely on late fusion also start with extraction of unimodal features. In contrast to early fusion, where features are then combined into a multimodal representation, approaches for late fusion learn semantic concepts directly from unimodal features. In [9] for example, separate generative probabilistic models are learned for the visual and textual modality. These scores are combined afterwards to yield a final detection score. In general, late fusion schemes combine learned unimodal scores into a multimodal representation. Then late fusion methods rely on supervised learning to classify semantic concepts. We define:

DEFINITION 2 (LATE FUSION). *Fusion scheme that first reduces unimodal features to separately learned concept scores, then these scores are integrated to learn concepts.*

Late fusion focuses on the individual strength of modalities. Unimodal concept detection scores are fused into a multimodal semantic representation rather than a feature representation. A big disadvantage of late fusion schemes is its expensiveness in terms of the learning effort, as every modality requires a separate supervised learning stage. Moreover, the combined representation requires an additional learning stage. Another disadvantage of the late fusion approach is the potential loss of correlation in mixed feature space. A general scheme for late fusion is illustrated in Figure 2.

3. FUSION SCHEME IMPLEMENTATION

We perform feature extraction on the visual and textual modality. After modality specific data processing, we combine features into a multimodal representation using an early fusion and late fusion scheme. The supervised learner is responsible for classifying the semantic concepts based on the feature patterns.

3.1 Visual Features Extraction

Visual feature extraction is based on the method described in [6]. In short, the procedure first extracts a number of invariant visual features per pixel. Based on these features the procedure labels each pixel in an image with one of 18 low-level visual concepts, like *concrete*, *sand*, *sky*, *water body*, and so on. This pixel-wise classification results in a labeled segmentation of an image f in terms of regional visual concepts. The percentage of pixels associated to each of the regional visual concepts is used as a visual content vector $\vec{w}_f$. To decide which f is the most representative for i, we select from all $\vec{w}_f$ in a shot the one that yields the highest score for a semantic concept. The feature vector $\vec{v}_i$, containing the best labeled segmentation, is the final result of the visual features extraction stage.

3.2 Textual Features Extraction

In the textual modality, we aim to learn the association between uttered speech and semantic concepts, see [6]. A detection system transcribes the speech into text. From the text we remove the frequently occurring stopwords. To learn the relation between uttered speech and concepts, we connect words to shots. We make this connection within the temporal boundaries of a shot. We derive a lexicon of uttered words that co-occur with ω using shot-based annotations from training data. For each concept ω, we learn a separate lexicon, Λ_ω, as this uttered word lexicon is specific for that concept. For feature extraction we compare the text associated with each shot with Λ_ω. This comparison yields a text vector $\vec{t}_i$ for shot i, which contains the histogram of the words in association with ω.

3.3 Supervised Learner

A large variety of supervised machine learning approaches exists to learn the relation between a concept ω and pattern x_i. For our purpose, the method of choice should handle typical problems related to semantic video analysis. Namely, it must learn from few examples, handle unbalanced data, and account for unknown or erroneous detected data. In such heavy demands, the Support Vector Machine (SVM)

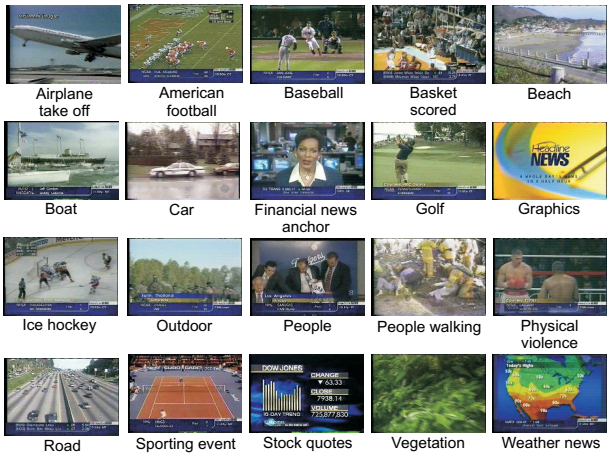

Figure 3: Instances of the 20 concepts in the lexicon.

framework [8] is a solid choice. We convert the SVM output using Platt's method [5] to acquire a measure in the form of a probability score. In addition, we perform parameter search on a large number of SVM parameter combinations to obtain good settings, $\vec{q}^*$ for the SVM. The result of the parameter search over $\vec{q}$ results in the model $p_i^*(\omega|x_i, \vec{q}^*)$ specific for ω.

3.4 Multimodal Features Combination

We rely on vector concatenation in both the early fusion and late fusion scheme to obtain a multimodal representation. We concatenate the visual vector $\vec{v}_i$ with the text vector $\vec{t}_i$. After feature normalization, we obtain early fusion vector $\vec{e}_i$. Then $\vec{e}_i$ serves as the input for an SVM, which learns a semantic concept for the early fusion scheme. For the late fusion scheme, we concatenate the probabilistic output score after visual analysis, i.e. $p_i^*(\omega|\vec{v}_i, \vec{q}^*)$, with the probabilistic score resulting from textual analysis, i.e. $p_i^*(\omega|\vec{t}_i, \vec{q}^*)$, into late fusion vector $\vec{l}_i$. Then $\vec{l}_i$ serves as the input for an SVM, which learns a semantic concept for the late fusion scheme.

4. EXPERIMENTAL SETUP

4.1 Data Set

We evaluate the early fusion and late fusion schemes within the TRECVID video retrieval benchmark [4]. The video archive of the 2004 TRECVID benchmark is composed of 184 hours of ABC World News Tonight and CNN Headline News. The development data contains approximately 120 hours. The test data contains the remaining 64 hours. Together with the video archive came automatic speech recognition results donated by LIMSI [2].

We split the 2004 TRECVID development data a priori into a non-overlapping training and validation set for our experiments. The training set $\mathcal{D}$ contained 85% of the development data, the validation set $\mathcal{V}$ contained the remaining 15%. We manually annotated a ground truth for all concepts considered. The semantic concepts in our lexicon are visualized in Figure 3. Together with the video data, the lexicons and annotated ground truth form the input for the sketched fusion schemes.

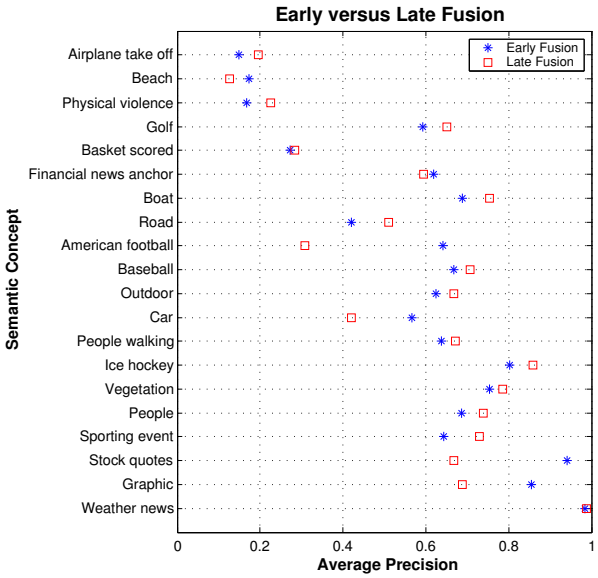

Figure 4: Comparison of early fusion versus late fusion for semantic indexing of 20 concepts.

4.2 Evaluation Criteria

We use *average precision* [4] to determine the accuracy of semantic concept detection at the shot level, following the standard in TRECVID evaluations. The average precision is a single-valued measure that corresponds to the area under a recall-precision curve.

We compose a pooled ground truth to reduce the labor-intensive manual judgments of all submitted runs. We take the pooled ground truth of TRECVID as a basis [4]. The top 100 results of both fusion schemes, for all 20 concepts, are then checked for presence in this pooled ground truth. Shots in the top 100 that have not been judged before are manually added to the pooled ground truth. We also add the ground truth for concepts that were not evaluated by TRECVID before. We then calculate average precision on this newly composed pooled ground truth. This pooling procedure allows for a fair comparison of our early and late fusion schemes.

5. RESULTS

We evaluated detection results for all 20 semantic concepts for both early fusion and late fusion. The results are visualized in Figure 4.

For the early fusion scheme, we trained the concepts on development set $\mathcal{D}$ only. We trained the unimodal features of the late fusion scheme on $\mathcal{D}$ also. Then we relied on $\mathcal{V}$ to make the final classification based on the multimodal combination. Due to this additional learning stage the late fusion scheme is able to obtain a better score for 14 concepts. The absolute difference ranges from 0.0 for *weather news* to 0.1 for *road*. More surprising is the result for the early fusion scheme, which obtains a better score for 6 concepts. Here, the absolute difference ranges from 0.0 for *weather news* to 0.3 for *stock quotes*. We conclude from these results that an additional learning stage doesn't necessarily have a positive effect on performance.

We plot the judged shots for three concepts in Figure 5,

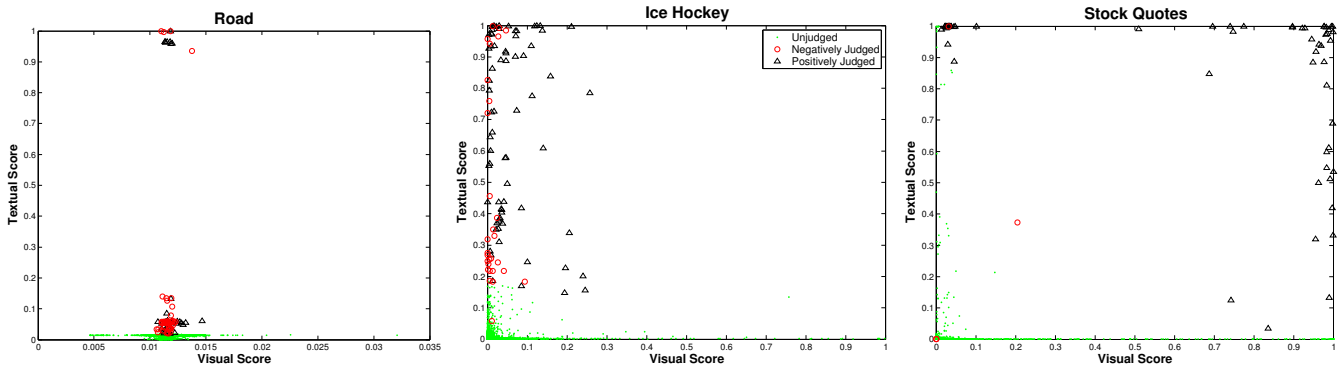

Figure 5: Distribution of judged shots for three semantic concepts using late fusion.

to gain insight in the fluctuating behavior of the late fusion scheme. For concepts *road* and *ice hockey*, the late fusion scheme is able to improve results. From the scores for the visual modality and the textual modality we observe that for the concepts *road* and *ice hockey* the scores either form a nice cluster (*road*) or are easily separable (*ice hockey*), see Figure 6 for *ice hockey* results. For *stock quotes* the situation is different. The late fusion scheme classifies a large number of easily separable scores correctly. But the late fusion scheme loses track for scores that are less prominent, resulting in a cluster of nearly 40 falsely judged shots that have a visual and textual score close to 0. This indicates that late fusion experiences difficulty in classifying shots that are close to the decision boundary of the SVM. These results suggest that a fusion strategy on a per-concept basis yields the most effective semantic index.

6. CONCLUSIONS

In this paper, we compare early fusion and late fusion schemes that aim to learn semantic concepts from multimodal video. Based on an experiment on 184 hours of broadcast video using 20 semantic concepts we conclude that a late fusion scheme tends to give better performance for most concepts, but it comes with the price of an increased learning effort. Moreover, if early fusion performs better the improvements are more significant. These results suggest that a fusion strategy on a per-concept basis yields an optimal strategy for semantic video analysis. We aim to evaluate such a hybrid fusion approach in future research.

7. ACKNOWLEDGMENTS

This research is sponsored by the BSIK MultimediaN project.

8. REFERENCES

[1] A. Amir et al. IBM research TRECVID-2003 video retrieval system. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2003.

[2] J. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1–2):89–108, 2002.

[3] G. Iyengar, H. Nock, and C. Neti. Discriminative model fusion for semantic concept detection and annotation in video. In *ACM Multimedia*, pages 255–258, Berkeley, USA, 2003.

[4] NIST. TREC Video Retrieval Evaluation, 2004. `http://www-nlpir.nist.gov/projects/trecvid/`.

[5] J. Platt. Probabilities for SV machines. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[6] C. Snoek et al. The MediaMill TRECVID 2004 semantic video search engine. In *Proc. TRECVID Workshop*, Gaithersburg, USA, 2004.

[7] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Trans. CSVT*, 11(4):522–535, 2001.

[8] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, USA, 2th edition, 2000.

[9] T. Westerveld et al. A probabilistic multimedia retrieval model and its evaluation. *EURASIP JASP*, (2):186–197, 2003.

[10] Y. Wu, E. Chang, K.-C. Chang, and J. Smith. Optimal multimodal fusion for multimedia data analysis. In *ACM Multimedia*, New York, USA, 2004.

Figure 6: Top 25 results for *ice hockey* using late fusion. Ordered left to right, top to bottom.