Interactive video retrieval based on multimodal dissimilarity representation

Eric Bruno ERIC.BRUNO@UNIGE.CH
Nicolas Moenne-Loccoz NICOLAS.MOENNE-LOCCOZ@UNIGE.CH
Stéphane Marchand-Maillet MARCHAND@UNIGE.CH
Viper group, Computer Vision and Multimedia Laboratory, University of Geneva
24 rue du Général Dufour, 1204 Geneva, Switzerland

Abstract

We present an approach to learn user semantic queries from dissimilarity representations of video audio-visual content. When dealing with large corpora of videos documents, using a feature-based representation calls for the online computation of distances between all documents and the query. Hence, a dissimilarity representation may be preferred because its offline computation speeds up the retrieval process. We show how distances related to visual and audio video features can directly be used to learn complex concepts from a set of positive and negative examples provided by the user. Based on the idea of dissimilarity spaces, we derive a low-dimensional multimodal representation space where an on-line and real-time classification is performed to learn user queries. The classification consists in maximizing a non-linear Fisher criterion to separate positive from negative examples. The evaluation, performed on the complete annotated TRECVid corpus, shows that our technique enables us to improve the precision of retrieval results.

1. Introduction

Determining semantic concepts by allowing users to iteratively refine their queries is a key issue in multimedia content-based retrieval. The relevance feedback loop allows to construct complex queries made out of positive and negative documents as examples. From this training set, a learning process should then extract relevant documents from feature spaces. Many relevance feedback techniques have been developed that operate directly in the feature space (Chang et al., 2003; Smith et al., 2003; Yan et al., 2003; Zhou & Huang, 2004).

Describing content of videos requires to deal in parallel with many high-dimensional feature spaces expressing the multimodal characteristics of the audiovisual stream. This mass of data makes retrieval operations computationally expensive when dealing directly with features. The simplest task of computing the distance between a query and all other elements becomes infeasible when involving tens of thousand of documents and thousand of feature space components. This problem is even more sensible when the similarity measures are complex functions or procedures, such as prediction functions for temporal distances (Bruno et al., 2005) or graph exploration for semantic similarities (Resnik, 1995).

A solution to allow on-line interaction would be to compute off-line monomodal dissimilarity relationships between elements and to use the dissimilarity matrices or distance-based indexing structures (Chávez et al., 2001) as an index for retrieval operations. The problem is then to find distance-based solutions that go beyond the classical k-NN approaches (Boldareva & Hiemstra, 2004) in order to perform effective classification and retrieval of semantic concepts. Pekalska *et al* (Pekalska et al., 2001) have proposed dissimilarity spaces where objects are represented not by their features but by their relative dissimilarities to a set of selected objects. These representations seem to form a convenient approach to tackle the similarity-based indexing and retrieval problem.

In this paper, we investigate the idea of dissimilarity spaces for the specific problem of multimedia document retrieval, and show how dissimilarities can be

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

used to build a low-dimensional multimodal representation space where learning machines based on *eg* non-linear discriminant analysis could operate. Our thorough evaluation on the complete TRECVid corpus shows that this multimodal dissimilarity space allows to perform effective retrieval of video documents in real time, as defined in (Nielsen, 1993).

2. Classification in dissimilarity space

In the proposed retrieval system, video segments are represented by their dissimilarity relationships computed offline over several audiovisual features. The user can formulate complex queries by iteratively providing positive and negatives examples in an online relevance feedback loop. From this training data, the aim is to perform a real-time dissimilarity-based classification that will return relevant documents to user.

2.1. Dissimilarity space

Let $d(\mathbf{x}_i, \mathbf{x}_j)$ be the distance between elements i and j according to their descriptors $\mathbf{x} \in \mathcal{F}$. $\mathcal{F}$ expresses the (unavailable) original feature space. The dissimilarity space is defined as the mapping $\mathbf{d}(\mathbf{z}, \Omega) : \mathcal{F} \rightarrow \mathbb{R}^N$ given by (see (Pekalska et al., 2001) for details):

$$\mathbf{d}(\mathbf{z}, \Omega) = [d(\mathbf{z}, \mathbf{x}_1), d(\mathbf{z}, \mathbf{x}_2), \dots\, d(\mathbf{z}, \mathbf{x}_N)]. \qquad (1)$$

The representation set $\Omega = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is a subset of N objects defining the new space. The new "features" of an input element are now the dissimilarities between itself and the representation objects. As a consequence, learning or classification tools for feature representations are also directly available to deal with the dissimilarities.

The dimensionality of the dissimilarity space is directly linked to the size of Ω, which controls the approximation made on the original feature space (such an approximation could be computed using projection algorithms like classical scaling (Cox & Cox, 1995)). Increasing the number of elements in Ω increases the representation accuracy. On the other hand, we are interested in minimizing the space dimensionality so as to limit computation and to speed up the response time of the system. The selection of Ω will however be driven by considerations on the classification problem as explained now.

2.2. Non-linear discriminant analysis

Let us define the set T as the query formed out of positive and negative training examples (respectively denoted $\mathcal{P}$ and $\mathcal{N}$ with $T = \mathcal{P} \cup \mathcal{N}$), their coordinates in the dissimilarity space are respectively $\mathbf{d}_i^+ = \mathbf{d}(\mathbf{z}_{i \in \mathcal{P}}, \Omega)$ and $\mathbf{d}_i^- = \mathbf{d}(\mathbf{z}_{i \in \mathcal{N}}, \Omega)$.

Given a query T, the aim is therefore to find a relevance measure $D(\mathbf{d}_i) : \mathbb{R}^N \rightarrow \mathbb{R}$ that maximizes the following Fisher criterion

$$\max_D \frac{\sum_i D^2(\mathbf{d}_i^-)}{\sum_i D^2(\mathbf{d}_i^+)}. \qquad (2)$$

The measure $D(\mathbf{d})$ gives us a new ranking function where positive elements tend to be placed at the top of the list while negatives one are pushed to the end.

Depending on the separability of the data according to a query T, the ranking function $D(\mathbf{d})$ may be chosen as a linear or non-linear function of the dissimilarities. Following the kernel machine formulation, $D(\mathbf{d})$ is written in both cases (linear or not) as an expansion of kernels centered on training patterns (Schölkopf & Smola, 2002):

$$D(\mathbf{d}) = \sum_{i \in T} \alpha_i k(\mathbf{d}, \mathbf{d}_i^{\pm}) + b. \qquad (3)$$

Using such non-linear model in criterion (2) leads to the formulation of the Kernel Fisher Discriminant (KFD) (Mika et al., 1999). It has been shown that this problem can be solved by using mathematical programs (quadratic or linear). The proofs and the implementation of the algorithm we use to optimize (2) can be found in (Mika et al., 2000).

In general, we are dealing with a $1 + x$ class setup with 1 class associated to positives and x to negatives (Zhou & Huang, 2004). It is then needed to estimate complex decision functions to learn the semantic concepts, increasing the risk to encounter difficulties for choosing and tuning well-adapted kernels. However, selecting the representation set as the set of positive examples $\mathcal{P}$ turns the problem into a binary classification. Assuming that the positive examples are close to each other while all being far from negatives, the vectors $\mathbf{d}(\mathbf{z}_{i \in \mathcal{P}}, \mathcal{P})$ (*within* scatter) have norms lower than vectors $\mathbf{d}(\mathbf{z}_{i \in \mathcal{N}}, \mathcal{P})$ (*between* scatter), leading to a binarization of the classification, as illustrated in figure 1. In addition, this choice readily induces to work in a low dimensional space of $p = |\mathcal{P}|$ components, where online learning processes are dramatically speeded-up.

Kernel selection and setting is a critical issue to successfully learn queries. It actually decides upon the classical trade-off between over-fitting and generalization properties of the classifier and hence is very dependent of the considered dissimilarity space. This problem is discussed in the next section.

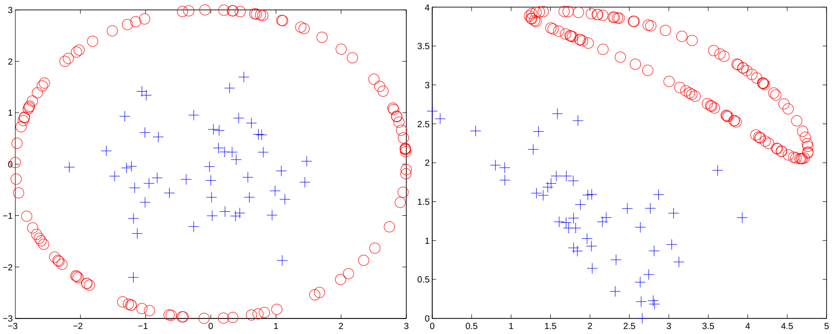

Figure 1. The $1 + x$ class problem in feature space (left) and dissimilarity space (right) where the representation objects are two points from the central class (cross)

3. Multimodal space

The video content is characterized by features corresponding to multiple modalities (*eg*, visual, audio, speech). Each of them leads to a dissimilarity matrix containing pairwise distances between all documents. Let us note d^{f_i} the distance measure applied on the feature space $\mathcal{F}_i$ and assume that dissimilarity matrices are known for M feature spaces. We define the multimodal dissimilarity space $\mathbf{d}$ as the concatenation of all monomodal spaces $\mathbf{d}^{f_i}$

$$\mathbf{d} = [\mathbf{d}^{f_1}, \mathbf{d}^{f_2}, \ldots, \mathbf{d}^{f_M}]. \qquad (4)$$

The kernel function used in equation (3) now operates in a multimodal space. Its choice is then a critical issue to ensure the success of the modalities fusion coming from the resolution of equation (2). The RBF kernel $k(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \mathbf{A} (\mathbf{x}-\mathbf{y})}$ presents a convenient solution for our problem: it is indeed able to learn semantic concepts that are locally distributed within the representation space, and the scaling symmetric positive definite matrix $\mathbf{A}$ permits to tune the trade-off between over-fitting and generalization. As the input space is multimodal, the scaling matrix is constructed so as to allow independent scaling for each feature space, so that $\mathbf{A} = \text{diag}[\boldsymbol{\sigma}_{f_1}, \cdots, \boldsymbol{\sigma}_{f_M}]$. The vector $\boldsymbol{\sigma}_{f_i} \in \mathbb{R}^p$ is constant with all values equal to the scale parameter σ_{f_i} estimated for the dissimilarity space $\mathbf{d}^{f_i}$. Various approaches to automatically tune the scale parameters (Cristianini et al., 2001; Ong et al., 2003) have been proposed. However, the kernel estimation rely on an optimization of functionals that will drastically penalize the response time of the retrieval system. For this reason, the estimation of σ_{f_i} is based on a less optimal but simpler heuristic, adapting the model to the query

$$\sigma_{f_i} = C \cdot \text{median}_i(\min_j ||\mathbf{d}_i^+ - \mathbf{d}_j^-||^2). \qquad (5)$$

In other words, the scale value in space $\mathbf{d}^{f_i}$ is set to be proportional to the median of all the minimum distances between the negative and the positive examples in that space. That way, the kernel becomes sharper as the two classes become closer to each other. The parameter C has been empirically set to 2.0.

4. Experimentations

Our multimodal interactive learning algorithm has been systematically experimented in the context of the video retrieval system we have developed. The segmented video documents, their multimodal description as well as manual annotations are stored in a database that keeps synchronized all data and allows large-scale evaluations of retrieval results.

The experimentation consists in making queries corresponding to annotated concepts and measuring the average precision (ratio of relevant documents in the retrieved list averaged over 50 queries) for retrieved lists of various lengths. The annotated positive examples are removed from the hitlist so that they are not taken into account when measuring the performance.

4.1. The video database

We use the complete annotated video corpus TRECVid-2003 composed of 133 hours of CNN and ABC news. Videos are segmented into shots and every shot has been annotated by several concepts. The speech transcripts extracted by Automatic Speech Recognition (ASR) at LIMSI laboratory (Gauvain et al., 2002) are also available.

We extracted the three following features from the 37'500 shots composing the corpus: Color histogram, Motion vector histogram and Word occurrence histogram (after stemming and stopping). The distance measures used are Euclidean for Color and Motion histogram and intersection for Word occurrence histogram.

4.2. Results

We first test the validity of the monomodal dissimilarity space defined in section 2.2. We compare the precision of the retrieval when the classification is performed in the color feature space and in the corresponding dissimilarity space. Figure 2 shows results for two queries corresponding to two annotated concepts (*Basketball* and *Studio setting*). Whatever the size of the training set, the precision at the 100th position of the retrieval list is better when the dissimilarity space is used. It is important to note that the improvement becomes more important when the training set is small: when the class distributions to estimate are severely under-sampled (small training set), the simplification of the classification problem implied by the dissimilarity space (see section 2.2) is crucial for the success of the training stage.

We now evaluate how the combination of modalities may improve the retrieval efficiency. Figure 3 compares the average precision for several concepts when the query is learned in the monomodal spaces and in the multimodal space. We can observe that, even for queries where the raw features used are not well-suited (*Car* and *Desert*), the combination of the three modalities performs better than considering them separately. The precision graphs also compare the algorithm with a random retrieval (e.g seeking hits at random within the database). This comparison illustrates the capability of the algorithm to use low-level multimodal information to create models of semantic concepts defined by user. This improves drastically the performance of the search.

The following experiment tests how the retrieval precision evolves when the number of positive and negative documents grows. As figure 4 shows, the precision of the retrieval increases with the size of the training set until a point where adding more examples does not improve the performances anymore. This behavior illustrates how the users, by providing more and more examples (relevance feedback loop), can refine their queries until reaching the optimum of the classifier.

Finally, since we act in an interactive setup, we were interested in the computation time problem. The following measures (table 1) have been done on a PIV 2GHz and include the time to access the dissimilarity matrices (37500×37500), the building of the multimodal dissimilarity space and the training of the Fisher classifier. As the dimensionality of the representation space linearly depends on the number of positive examples, the response time increases according to their number. On the other hand, negative examples have less influence since they are just involved in the learning process.

Table 1. Response time

Neg. examples	20			100
Pos. examples	5	10	40	10
Resp. time (s)	1.4	2	7.4	4.3

5. Conclusion

We have presented a retrieval strategy for video documents. Based on a multimodal dissimilarity space associated to a non-linear discriminant analysis, the algorithm is able to take benefit from low-level multimodal descriptions of video documents and, as a consequence, to learn semantic queries from a limited number of input examples. The design of the dissimilarity space has been achieved so as to simplify the classification problem while building a low-dimensional representation of the data. The use of the positives examples as a representation set transforms the $1 + x$ setup into a binary classification problem. Sophisticated learning machines, such as the kernel Fisher discriminant analysis, can then directly be applied to classify data. As a result, semantic concepts are learned with more efficiency and queries on large databases are processed near real-time which authorizes the use of feedback loop as a search paradigm. Extensive evaluations on the TRECVid-2003 benchmark show the efficiency and the usability of the proposed multimodal space and fusion algorithm to retrieve documents within a large corpus of videos.

While the presented classification scheme has proved its value, the actual features considered to characterize the videos do not permit us to design a fully-capable and efficient video retrieval system. The design of new feature extractors related to new modalities (*e.g.* audio stream) and higher-level aspects of the content (*e.g.* face and object detection) is still a major issue. The addition of information sources should leads us to investigate more deeply the problems of the multimodal kernel design and setting as well as to determine the limits of the fusion scheme when a large number of features is used.

References

Boldareva, L., & Hiemstra, D. (2004). Interactive content-based retrieval using pre-computed object-object similarities. *Conference on Image and Video Retrieval, CIVR'04* (pp. 308–316). Dublin, Ireland.

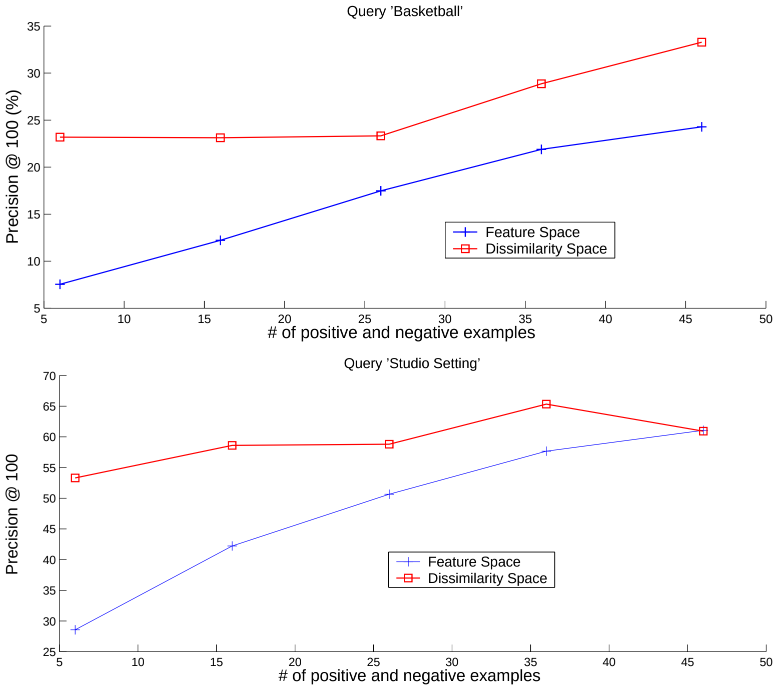

Figure 2. Average precision of the retrieval at 100 when the classification is performed directly in the color feature space (cross) and in the corresponding dissimilarity space (square) as the number of positive and negative examples increases.

Bruno, E., Moenne-Loccoz, N., & Marchand-Maillet, S. (2005). Unsupervised event discrimination based on nonlinear temporal modelling of activity. *Pattern Analysis and Application, special issue on Video Event Mining*. (to appear).

Chang, E. Y., Li, B., Wu, G., & Go, K. (2003). Statistical learning for effective visual information retrieval. *Proceedings of the IEEE International Conference on Image Processing.*

Chávez, E., Navarro, G., Baeza-Yates, R., & Marroquin, J. (2001). Searching in metric spaces. *ACM Computing Surveys*, *33*, 273–321.

Cox, T., & Cox, M. (1995). *Multidimensional scaling.* London: Chapman & Hall.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2001). On kernel-target alignment. *Advances In Neural Information Processing Systems, Nips.*

Gauvain, J., Lamel, L., & Adda, G. (2002). The limsi broadcast news transcription system. *Speech Communication*, *37*, 89–108.

Mika, S., Rätsch, G., & Müller, K.-R. (2000). A mathematical programming approach to the kernel fisher algorithm. *NIPS* (pp. 591–597).

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX* (pp. 41–48). IEEE.

Nielsen, J. (1993). *Usability engineering.* Boston, MA, USA: Academic Press.

Ong, C., Smola, A., & Williamson, R. (2003). Hyperkernels. *Advances in Neural Information Processing Systems, NIPS.*

Pekalska, E., Paclík, P., & Duin, R. (2001). A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, *2*, 175–211.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *14th International Joint Conference on Artificial Intelligence, IJCAI* (pp. 448–453). Montreal, Canada.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels.* MIT Press.

Smith, J. R., Jaimes, A., Lin, C.-Y., Naphade, M., Natsev, A., & Tseng, B. (2003). Interactive search fusion methods for video database retrieval. *IEEE International Conference on Image Processing (ICIP).*

Yan, R., Hauptmann, A., & Jin, R. (2003). Negative pseudo-relevance feedback in content-based video retrieval. *Proceedings of ACM Multimedia (MM2003).* Berkeley, USA.

Zhou, X., & Huang, T. (2004). Small sample learning during multimedia retrieval using biasmap. *Proceedings of the IEEE Conference on Pattern Recognition and Computer Vision, CVPR'01* (pp. 11–17). Hawaii.

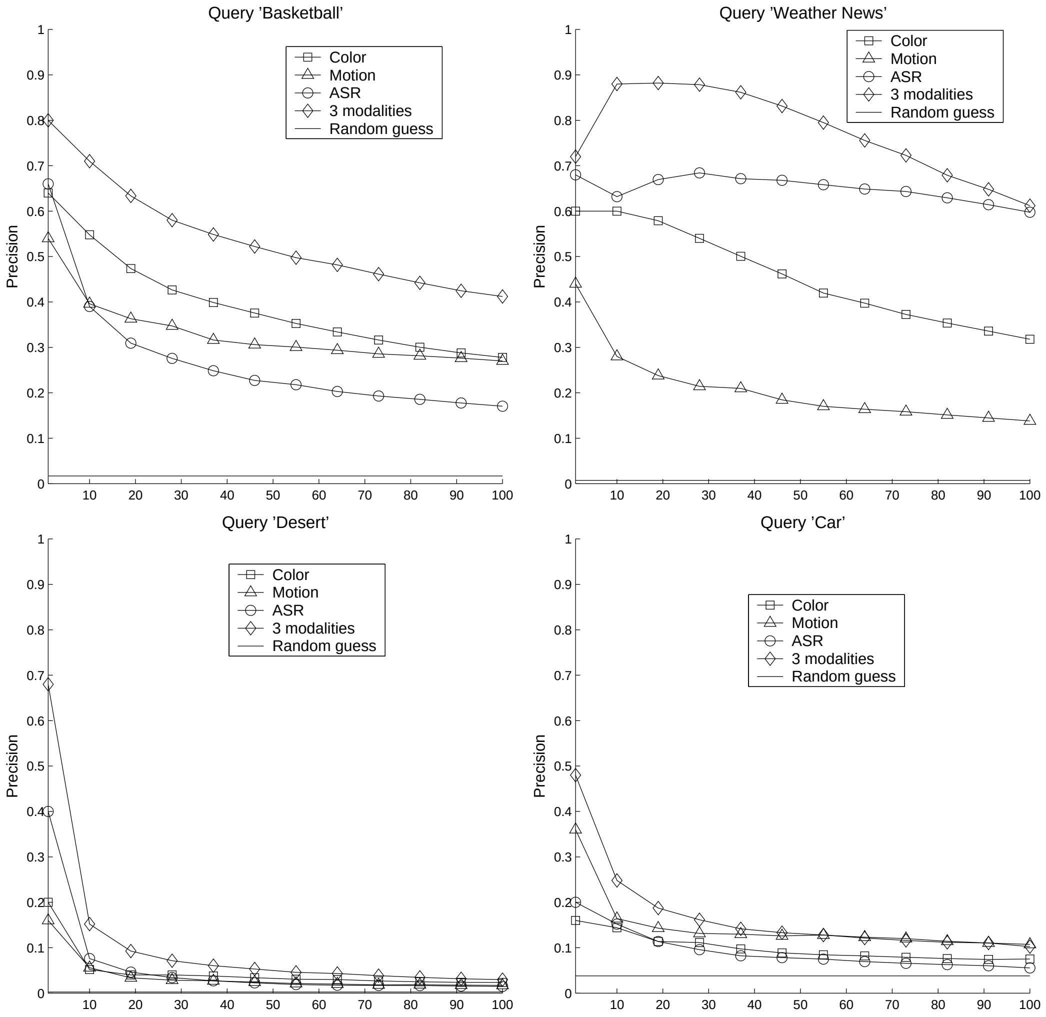

Figure 3. Average precision vs. length of retrieved lists for monomodal and multimodal dissimilarity spaces. The query is composed of 5 positive examples (annotated by the concept) and 20 negative examples randomly selected in the database. The "random guess" line is equal to the proportion of the concept in the database.

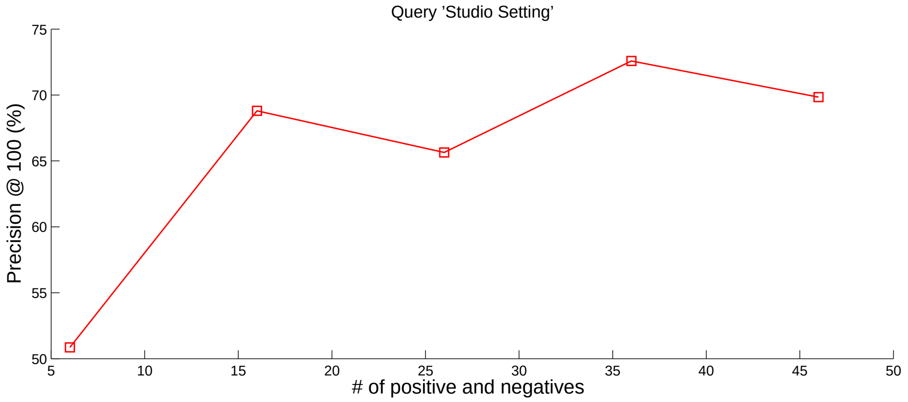

Figure 4. Average precision at 100 when positive examples and negative examples increase ($n_p = n_n$). Color, motion and ASR features are used.