A CONTEXTUAL MODEL FOR SEMANTIC VIDEO STRUCTURING

Bruno Janvier, Eric Bruno, Stephane Marchand-Maillet, Thierry Pun

Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva
24 rue Général Dufour, CH-1211, Geneva 4, Switzerland
phone: + (41) (0)22 3797147, email: janvier@cui.unige.ch
web: http://vision.unige.ch

ABSTRACT

The problem of semantic video structuring is vital for automated management of large video collections. The goal is to automatically extract from the raw data the inner structure of a video collection ; so that a whole new range of applications to browse and search video collections can be derived out of this high-level segmentation. To reach this goal, we exploit techniques that consider the full spectrum of video content ; it is fundamental to properly integrate technologies from the fields of computer vision, audio analysis, natural language processing and machine learning.

In this paper, a multimodal feature vector providing a rich description of the audio, visual and text modalities is first constructed. Boosted Random Fields are then used to learn two types of relationships : between features and labels and between labels associated with various modalities for improved consistency of the results. The parameters of this enhanced model are found iteratively by using two successive stages of Boosting.

We experimented using the TRECvid corpus and show results that validate the approach over existing studies.

1. INTRODUCTION

Large archives of broadcast news video have already been collected in digital form and are rapidly growing in quantity. Documentalists need effective tools for the management of video collections. This paper presents an approach towards semantic-level automated partitionning to segment raw MPEG videos into semantically meaningful story units as defined as "'news story segmentation"' task in the TRECVID 2003 experiment [1].

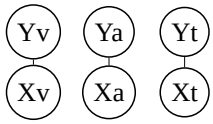
The challenge of news story segmentation is to estimate the probability $p(b|x)$ that there is a story boundary b at a given position in the video stream knowing a context x which contains information about various multimodal sources : visual, audio and text information coming from the video and audio streams of MPEG files and Automatic Speech Recognition (ASR) transcripts. The researcher in this subject has to integrate technologies from the fields of computer vision, audio analysis and natural language processing as well as machine learning to reach his goal which is to allow facilitated access to multimedia data.

However, this problem is very ill-defined. Many cues are possibly useful, but none of them are self-sufficient to make an optimal decision. We assert that by combining cues from every possible modality (visual, audio and textual), it will be possible for the system to obtain a correct decision.

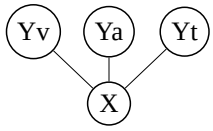
The first step is to build a large pool of features from each modality for the low-level description of the video content. Then we distinguish three different approaches for such classification problems as shown in the figure 1. The classical approach consists in using visual features X_v to infer a visual class Y_v and independently using audio features X_a to infer an audio class Y_a and so on. Such approaches are still commonly used by research groups that belong to a community focusing on one particular modality.

The multimodal approach is about using a combination of features X from all modalities to infer visual, audio and text labels. As an example, the word 'temperature' which a textual feature is likely to be useful to infer the visual label 'weather news'. Multimodal classification is generally done using SVMs or Boosting. For example in [5], several multimodal classification methods are compared for the purpose of news story segmentation.

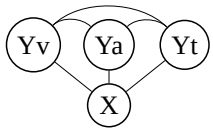
The contextual and multimodal approach goes further by considering not only the link between feature vectors and labels but also the relationships between labels concerning different modalities : in our case, a context is thus defined as the compatibility for a label to appear together with labels concerning other modalities. As an example, if the multimodal classification infered that the visual label is likely to be 'news subject monologue', the model will use this information to infer the label for the audio content which is more likely to be 'speech'.



Classical classification



Multimodal classification



Contextual and multimodal classification

Figure 1: Various classification models where $X = (Xv, Xa, Xt)$.

For every video segment, we have to deal altogether with a visual, audio and textual classification problem. For every

boundary separing two consecutive video segments, we also have to deal with a binary classification to tell whether there is is a news story boundary or not.

An extended use of the context is the key to enable accurate video content modeling as accurately as possible. We propose to build a contextual model designed for news story segmentation and to use Boosted Random Fields, as presented by A. Torralba *et al.* in [11], to estimate the parameters of the model. Boosted random fields share the same ability as boosting to dig deeply into the set of features to select the ones which are really helpful to improve the classification results without overfitting. It is also a promising new approach to model relationships between labels so that higher consistency of the classification results can be achieved.

2. CANDIDATE POINTS SELECTION

We will assign multimodal labels to all shot units. In [3], we proposed an algorithm to decompose a video into color homogeneous segments. We showed that the approach was robust to detect all kinds of transitions between different shots in a generic manner.

A simple and useful cue to detect a news story boundary is the presence of a short silence during the transition from one shot to another, as the anchor persons usually mark a short pause when switching from one story to the next.

We will use this heuristic as necessary condition for a story boundary to be present. Hencec, the candidate points will be the union of the set of a boundaries between shots and the set of audio silences with a tolerance window of one second.

3. DESIGN OF AN EXPRESSIVE FEATURE POOL

The first issue to solve is to provide our classifier with an expressive feature pool so that the algorithm might find relationships between features and labels.

In the context of the TRECVID experiment, we have at our disposal annotations for the following semantic classes in the training set:

- visual content: news subject monologue, studio-settings, outdoors, man-made scene, cartoon, weather news, sport scene, text scene, graphics, ads
- audio content: speech, music, noise, speech+music, speech+noise, other sound

For the text content, we do not have a set of predefined and annotated classes. We will hence use an unsupervised clustering algorithm to define a set of N classes from the training data.

We have chosen to use the Information Bottleneck algorithm [2] which has proven to be powerful to cluster similar topics together.

The pool of features has to be chosen in order to have a high discriminative power according to these set of labels. The goal here is to feed the system with a representation of features describing the video content for semantic classification as well as news story boundary detection. Another restriction is that the set of features has to be generic enough to adapt to different datasets like CNN or ABC news broadcasts.

3.1 Visual information

For a given video segment, the color information will be represented by a color histogram of the keyframe. This is needed to characterize particular backgrounds in frequently occuring video segments.

The motion information will be an histogram of the horizontal and vertical components of the optical flow directly extracted from the MPEG stream. A measure of the motion intensity will also be added to distinghish between small, medium and high level of actions. Motion is particularly important to discriminate sports scene/ads from news subject monologue/weather news as an example.

3.2 Audio information

The features used to represent the audio information will be computed on one second clips containing frames of 50ms. We will use a perceptual and expressive set of features : the spectrum and cepstrum flux (SF, CF) which measures the average variation frame per frames of the spectrum or cepstrum, the bandwidth (BW) which measures the repartition of frequencies around its center of gravity, the energy ratio of a subband to the total energy (ERSB) where the audio spectrum is decomposed into 4 perceptual subbands from $0-630-1720-4400-inf$ Hz and the energy ratios are computed, the frequency component volume contour at $4Hz$ (FCVC4) where we multiply the spectrum by a triangular window around 4 Hz and calculate the energy (it is particularly appropriate to detect speech), the low energy ratio(LSTER) which measures the number of frames with low energy ($< 0.5 * mean$) in the clip, the spectral centroid (SC), the spectral rolloff (SR) which is another measure of the repartition of the spectrum (it is the point below which 85% of the magnitude lies), the root mean square energy (RMS) which measures the mean volume of the clip, the volume standard deviation (VSTD) which measures the repartition of the volume during the clip, the zero crossing rate (ZCR) which counts the number of times the audio has crossed 0, the high zero crossing rate ratio (HZCRR) where the ratio of frames with a high zero crossing ($> 1.5 * \text{mean}$) rates in the window.

3.3 Text information

The text information is provided in the TRECVID evaluation data in the form of time-stamped ASR transcripts. The ASR engine has been developped by the LIMSI [8]. The text content is modeled by the word vector representation after stemming and stop-words removal. A vector is also composed to denote the presence/absence of a list of N-grams relevant to news story transitions and which has been constructed previously.

4. CONTEXTUAL AND MULTIMODAL CLASSIFICATION

Multimodal classification approaches have been used to solve the news story segmentation within the TRECVID community. Hsu *et al.* [4] have used a maximum entropy model [6] to fuse binary, multimodal cues. The interest of using a maximum entropy classifier is that features can be mutually dependent. It does perform better than a Bayesian classifier for such kind of problems. In [5], the same author demonstrates using a number of experiments that news story

segmentation can be achieved using various techniques for multimodal classification : maximum entropy, boosting and support vector machines. In Hsu *et al.* experiment, support vector machines show superior performances. However, the features and classification are local and no context is taken into account. Chaisorn *et al.* [7] have introduced the idea of a two-level multimodal system. The semantic labelling of video shots classification is first performed by using decision trees. Then a Hidden Markov Model (HMM) is trained using the labels as observations for the segmentation in news stories. We propose a richer representation of the context because every modality will be mapped onto a different label alphabet.

4.1 Semantic labelling of video segments and News story segmentation

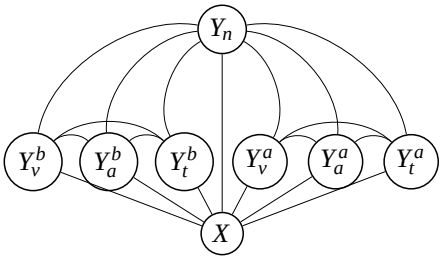


Figure 2: Contextual News story segmentation. The subscripts *n,v,a,t* denotes 'news boundary','visual', 'audio','text' labels.The superscripts *b* and *a* denotes video segments located 'before' and 'after' the candidate points.

As shown in the figure 2, we wish to classify jointly the set of candidate points with the news boundary/non-news boundary labels as well as the video segments located before and after into a set of semantic labels concerning the three modalities. For the video segment before/after the candidate point, the classification should take into account the interaction between labels to improve consistency. For example, it will reduce the possibilities to see the visual modality be assigned the label 'sports' if audio and text are pointing toward 'financial news'. Also, in this model, all labels related to all 3 modalities before and after will interact with the news story segmentation labels. Finally and of course, all labels are interacting with our observable and low-level features set. At this stage, we face a very complex model taking into account important relationships.

Boosted Random Fields gives the theoretical tools and optimization framework to estimate the parameters of such a complex model in a consistent way. In the following, X is a random variable over data to be labeled and Y are random variables over corresponding visual, audio, textual and news boundary labels with respect to different finite labels alphabet. For each node i, N_i corresponds to the neighbors of the node i. The distribution of the labels conditionned on X is the same as for Conditional Random Fields [9] and is given by :

$$p(Y|X) = \frac{1}{Z}\prod_i \left(\phi_i(Y_i,X) \prod_{j\in N_i} \psi_{i,j}(Y_i,Y_j,X) \right)$$

where the functions ϕ_i corresponds to evidence associating locally observed data and labels. The $\psi_{i,j}$ are compatibility potentials depending on observed data and also on neighboring labels, allowing label interaction. Z is the partition function.

With such a model, the problem of learning the ϕ_i, even if the functions $\psi_{i,j}$ are known, is untractable when the graph structure is not a chain or a tree because of the computation of Z (an exponential sum). Boosted Random Fields solves this problem by constructing an iterative approximation of the solution by using an additive model for the local evidence and label compatibility functions.

The training data contains M instances of training pairs (x_{im}, y_{im}). In the rest of this explanation, only binary classes will be considered. For a larger number of classes L, we will proceed as for the AdaBoost.MH algorithm defined by Shapire and Singer : by expanding the M observations into $M \times L$ pairs $((x_{im},1),y_{im1}),...,(x_{im},L),y_{imL}))$ where $m = 1,...,M$ and $y_{iml} \in \{-1,1\}$ response for node i, observation m and class l.

The cost function to minimize at iteration t is the per-label loss :

$$J^t = -\prod_m \prod_i p(y_{im}=+1|x_{im},t)^{\frac{y_{im}+1}{2}} p(y_{im}=-1|x_{im},t)^{1-\frac{y_{im}+1}{2}}$$

where the belief $p(y_i|x_i,t)$ of a node i at iteration t is proportional to the local evidence function multiplied by the product of all messages coming into the node i from all of its neighbors.

$$p(y_i|x_i,t) \propto \phi_i^t(y_i) Mess_i^t(y_i)$$

where $Mess_i^t$ is the product of all messages coming from neighbors which are function of the belief and of the compatibility functions ψ:

$$\begin{aligned} Mess_i^{t+1}(y_i) &= \prod_{k\in N_i} \mu_{k\to i}^{t+1}(y_i) \\ \mu_{k\to i}^{t+1}(y_i) &= \sum_{y_k\in\{-1,+1\}} \psi_{k,i}(y_k,y_i) \frac{p(y_k|x_k,t)}{\mu_{i\to k}^t(y_k)} \end{aligned}$$

If we denote :

$$\begin{aligned} F_i^t &= \frac{log(\phi_i^t)}{y_i} \\ G_i^t &= \log M_i^t(+1) - \log M_i^t(-1) \\ p(y_i|x_i,t) &= \frac{1}{1+e^{-(F_i^t+G_i^t)}} \end{aligned}$$

F_i^t and G_i^t are direct function of the local evidence and of the compatibility potentials respectively. It can be shown that the cost function simplifies to :

$$\log J_i^t = \sum_m \log(1+e^{-Y_i(F_i^t+G_i^t))})$$

The main idea is not to estimate the functions ϕ_i and $\psi_{i,j}$ directly, but to minimize the cost function iteratively via two successive stages of Boosting by using an additive model for F_i^t and G_i^t.

$$\begin{aligned} F_{i,m}^t &= \sum_{n=1}^t f_i^n(x_{i,m}) \\ G_{i,m}^t &= \sum_{n=1}^t g_i^n(b_m^t) \end{aligned}$$

The functions f_i^n and g_i^n are weak learners in the form of regression stumps. The functions f_i^n are dependent on the features $x_{i,m}$ of the training data whereas the functions g_i^n are dependent on the beliefs b_m^t. Iteratively, the weak learners f_i^n and g_i^n are chosen so that the cost function is minimized ; this is done by weighted least square :

$$\text{argmin} \log(J_i^t) \quad = \quad \text{argmin} \sum_m w_m^t (Y_m^t - f_i^t(x_{i,m}))^2$$

The time complexity of the algorithm is a linear function of the number of iterations T, the number of nodes N and the number of features V and is written O(T.N.V).

5. EXPERIMENTS AND DISCUSSION

5.1 Training and test data

The training and test sets contain video files from CNN broadcast news from the TRECVid corpus. 50 percent of the data was used for training and 50 percent for testing. The CNN collection contains 34 videos and every video is made of an average of 400 candidate points.

5.2 Semantic video segment labelling

Speech	Music	Noise
78	64	32

Table 1: Percent of correct classification for the audio labels

Anchor person	Outdoor	Financial	Weather	Ads
67	53	71	83	55

Table 2: Percent of correct classification for the visual labels

Tables 1 and 2 highlight the fact that varying performances are obtained when considering different labels.

5.3 News story segmentation

The measure of performance for the semantic segmentation is done by calculating the precision and recall of the system. Every reference boundary is enlarged by a tolerance window of 5s in both directions according to the TRECvid evaluation protocol.

Non-contextual

P	R
0.57	0.62

Contextual

P	R
0.64	0.66

Precision and recall for the news story segmentation

The use of context shows a clear improvement in performance compared with the regular adaboost which classifies the data without considering their compatibilities. These performances may be improved further by selecting the set of labels which are the more informative to help the decision process. Filtering the labels which offers a poor classification accuracy should also help furthere.

6. CONCLUSION

In this work, we have proposed a contextual model for the semantic segmentation of videos into stories by allowing labels interactions between different modalities as well as the observed data. Boosted Random Fields provide a principled approach and an effective optimization framework to estimate the model parameters. The results on the TRECVid corpus validate the advantages of this model. In the future, we intend to explore different ways to find the most discriminative labels so that the context can be defined in a manner which is maximally informative. It would also be interesting to use such models for the high-level feature extraction task for semantic labelling of video content as defined in the context of the TRECVid experiment.

7. AKNOWLEDGMENTS

This work is supported by the Swiss National Center of Competence IM2 - Interactive Multimedia Information Management.

REFERENCES

[1] *http://www-nlpir.nist.gov/projects/trecvid/*

[2] G. Chechik. and N. Tishby, Extracting relevant structures with side information, *Advances in Neural Information Proceedings Systems NIPS*, 2002

[3] B. Janvier and E. Bruno and S. Marchand-Maillet and T. Pun, Information-Theoretic Framework for The Joint Temporal Partionning and Representation of Video Data, *Proceedings of the 3rd International Workshop on Content-Based Multimedia Indexing*, CBMI'03, Rennes, France, September 2003

[4] W. Hsu and S.F. Chang and C.W Huang and L. Kennedy and C.Y. Lin and G. Iyengar, Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation, *SPIE Electronic Imaging*, 2004

[5] W. Hsu and S.F. Chang, Generative, discriminative and ensemble learning on multi-modal perceptual fusion toward news video story segmentation, *IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, 2004

[6] D. Beeferman and A. Berger and J. Lafferty, Statisticals models for text segmentation, *Machine Learning 34 (special issue on Natural Language Learning)*, 177-210, 1999

[7] L. Chaisorn and T-S. Chua and C-K. Koh and Y. Zhao and H. Xu and H. Fend and Q. Tian, A two-level multimodal approach for story segmentation of large news video corpus, *TrecVid Workshop)*, 2003

[8] J.L. Gauvain, L. Lamel, and G. Adda., The LIMSI Broadcast News Transcription System., *Speech Communication*, 37(1-2):89-108, 2002

[9] J. Lafferty and A. McCallum and F. Pereira, Conditional random fields : probabilistic models for segmenting and labeling sequence data, *Proc. 18th International Conf. on Machine Learning*, 2001

[10] R. Shapire and Y. Singer, Improved boosting algorithms using confidence-rated predictions, in *Proceeding of the Eleventh Annual Conference on Computational Learning Theory*, 1998

[11] A. Torralba and K. P. Murphy and W. T. Freeman, Contextual models for object detection using boosted random fields, *NIPS*,2004