Semantic Segmentation of Video Collections using Boosted Random Fields

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

Université de Genève, 24 rue Général Dufour, CH-1211 Switzerland,
janvier@cui.unige.ch

Abstract. Multimedia documentalists need effective tools to organize and search into large video collections. Semantic video structuring consists in automatically extracting from the raw data the inner structure of a video collection. This high-level information if automatically extracted would provide important meta information enabling the development of an important new range of applications to browse and search video collections.
In this paper, we present the feature extraction process providing a compact description of the audio, visual and text modalities. To reach the semantic level required, a contextual model is then proposed: it is a complex model which takes into account not only the link between features and labels but also the compatibility between labels associated with different modalities for improved consistency of the results. Boosted Random Fields are used to learn these relationships. It provides an iterative optimization framework to learn the model parameters and uses the abilities of boosting to reduce classification errors, to avoid over-fitting and to achieve the task of feature selection.
We experiment using the TRECvid corpus and show results that validate the approach over existing studies.

1 Introduction

The amount of digital archives of broadcast news videos is quickly growing in quantity. Multimedia documentalists need effective tools for the management of video collections. This article presents an approach towards automated semantic partitionning to segment raw MPEG videos into semantically meaningful story units and focus on the learning task. Our evaluation is based on a collection of broadcast news videos used for the TRECVID 2003 experiment [1].

News story segmentation is about the estimation of the probability $p(b|x)$ that there exists 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. It has to deal with the semantic gap problem: the lack of coincidence between the formative and cognitive information. It is quite ill-defined. Many cues are possibly useful, but none of them are self-sufficient to make an optimal decision: this is by the

combination of multimodal cues (visual, audio and textual) to infer multimodal labels and by considering the interactions between these multimodal labels that we assert the system will manage to make correct decisions. The researcher in this subject has to integrate results from the fields of computer vision, audio analysis and natural language processing as well as machine learning to reach the goal of allowing facilitated access to multimedia data.

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 coming from every modality 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. In this article [6], 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 associated with different modalities : in our case, a context is thus defined as the compatibility for a label to appear together with labels related to other modalities. For 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'. Another example is if the visual label is infered to be 'Weather news' with low confidence whereas the text and audio labels infered are highly incompatible with 'Weather news', it is likely that the final decision will not be 'Weather news' but another label which is more compatible with the different modalities according to the model.

An extended use of the context is the key to be able to enable accurate video 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 [12], to estimate the parameters of the model. We will focus this discussion on the advantage of using Boosted random fields. It 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.

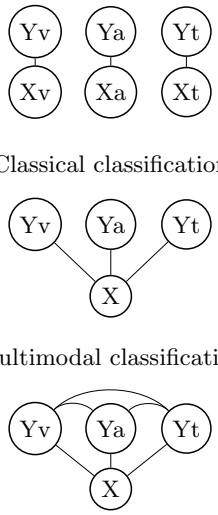

Fig. 1. Various classification models where $X = (Xv, Xa, Xt)$.

2 Candidate points

We will not assign multimodal labels to all shot units. We will restrict the search for news story boundaries to a reduced set of candidate points and only considers the labels of the shots which are located around them.

In [4], we have proposed an algorithm to decompose a video into color homogeneous segments. We showed that the approach was robust to detect all types of transitions between different shots in a generic manner. By using shot boundaries alone as candidate points, the set covers 91% of the news story boundaries.

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. By combining with silence detection, the set of candidate points is reduced by 33% and cover 87% of the news story boundary. Hence the candidate points will be the union of the set of shot boundaries and the set of audio silences with a tolerance window of one second.

3 Labels selection

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 any set of predefined and annotated classes. We will therefore 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] [3] which has proven to be powerful to cluster similar topics together. Given the joint distribution $p(N, V)$ of news stories (N) and vocabulary (V), the information bottleneck is searching for a compact set of news clusters (T) which preserve information about the variable V. The mutual information is used as an optimal metric from the information theory point of view. The goal is to compress the set of news stories while preserving as much information as possible about the distribution of words. Hence the functional to minimize is given by the following trade-off : $F = I(T;V) - \frac{1}{\beta} I(T;N)$. The trade-off β can be set to infinity if we are only interested to maximize relevant information V only. As β will get closer to 1, less balanced clustering solutions will be found where the small clusters T are found so that there is a minimal loss of information about N. In the experiment, we chose to reduce the set of textual labels to 20 and we did set the parameter β to the value 50.

	Most frequent words
Cluster 1	rain weather with storm continue today across temperature forecast new california coast
Cluster 2	med won gold olymp when team with game five two today hi second sport
Cluster 3	presid clinton today south with say hi africa nate first visit lead talk
Cluster 4	day with look go what plain like all wait make get again off well first win season
Cluster 5	presid with house clinton white about lawyer investigation jury grand said case sexual
Cluster 6	point nasdaq dow wall street gain back market stock industrial today jones eighty seven close
...	...

Table 1. News story clustering into a set of 20 topics

The table 1 shows a short summary of the content of the different news story clusters found thanks to the information bottleneck algorithm. We will use the cluster numbers as labels for the textual modality.

4 Multimodal feature extraction

An important issue is to provide our classifier with an expressive feature pool so that the algorithm might find relationships between features and labels. We need here to obtain 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 containing various news broadcasts (CNN, ABC, ...).

4.1 Visual information

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

The motion information is given as an histogram of the horizontal and vertical components of the optical flow directly extracted from the MPEG stream. A measure of the motion intensity is also 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 for example.

4.2 Audio information

The features used to represent the audio information will be computed on one-second clips. They are generally measured by considering FFT calculations over frames of 50ms. We use a perceptual and expressive set of features:

- the spectrum and cepstrum flux (SF, CF) which measures the average frame per frame variation 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 $0-630-1720-4400-\text{inf}$ 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 is)
- 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

4.3 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 [9]. The text content is modeled by the word vector representation after stemming and stop-words removal. A vector is also formed to denote the presence/absence of a list of N-grams relevant to news story transitions and which has been constructed previously.

5 Contextual model description and learning algorithm

5.1 Related work

Multimodal classification approaches have been used to solve the news story segmentation problem within the TRECVID community. Hsu *et al.* [5] have used a maximum entropy model [7] to fuse binary multimodal cues. The interest of using a maximum entropy classifier is that features can be mutually dependent. It performs then better than a Bayesian classifier for such kind of problems. In [6], the same author demonstrates using a number of experiments that news story segmentation can be achieved with 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.* [8] 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. Here, the representation of the context we propose is richer since every modality will be mapped onto a different label alphabet.

5.2 Description of the contextual model

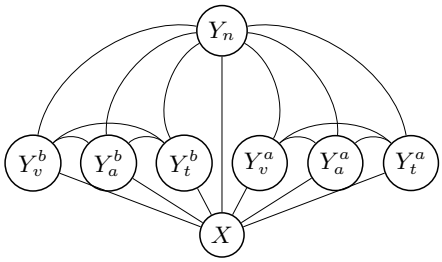

Fig. 2. Contextual News story segmentation. The subscripts n,v,a,t denotes 'news boundary','visual', 'audio','text' labels.The superscripts b and a denotes video segments located 'before' and 'after' the candidate points.

As shown in the figure 2, we wish to jointly classify the set of candidate points with the news boundary/non-news boundary labels as well as the video segments located before and after the considered candidate point into a set of semantic labels related to the three modalities. For the video segment before/after the candidate points, 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 both 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 Y_n. Finally and of course, all labels are interacting with our observable and low-level features set X. At this stage, we face a very complex model taking into account: the multimodality of features, the relationships between modalities for the labels and temporal modeling.

5.3 Classification using Boosted Random Fields

Boosted Random Fields provide 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 [10] 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 neighbor labels, allowing label interaction. Z is the partition function. It is possible to note at this stage that the model is discriminative: the focus is on the conditional probability $p(Y|X)$ and we will never consider the joint probability $p(Y, X)$ as it is done when using generative models such as hidden Markov models (HMMs). However we still provide a temporal modeling of the video data by considering shots located before and after the candidate points.

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 solve 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 [11] : 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 defined by:

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

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 :

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

5.4 Discussion

The training algorithm is basically for a given number of iterations:

- search for the optimal weak learner on the features to calculate f
- search for the optimal weak learner on the beliefs to calculate g
- update according to the equations F, G
- update $p(y_i|x_i,t)$ and $w = p(1|x_i,t)p(-1|x_i,t)$

The two stages of boosting are combined to jointly converge to a solution that takes into account local potentials F and compatibilities between labels G. As in boosting, the reweighting by w of the training instances pushes the algorithm to focus on hard examples. When estimating f and g, the feature and label selection is done by the weak learners. All features and labels are considered, but only the ones which are useful to reduce the classification errors are selected.

It is similar to a two-stage learning algorithm where the labels learned at the first stage are then used as features for the second stage commonly used to deal with the semantic gap (see [8] for a combination of decision trees and HMM) with the major difference that all labels are learned jointly and interact together during the learning process.

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 $O(T.N.V)$.

6 Experiments

6.1 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.

6.2 Semantic video segment labelling

Speech	Music	Noise
78	64	32

Fig. 3. Percent of correct classification for the audio labels

Anchor	Outdoor	Financial	Weather	Ads
67	53	71	83	55

Fig. 4. Percent of correct classification for the visual labels

Figures 3 and 4 highlight the fact that varying performances are obtained when considering different labels.

6.3 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

Fig. 5. Precision and Recall for the news story segmentation with a contextual or non-contextual model

The use of context shows a clear improvement in performance compared with the regular adaboost which classifies the data without considering the compatibilities of the labels. 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 further.

7 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. The next step is to provide more discriminative features to the learning algorithm. In the future, we intend to explore different ways 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.

8 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. N.Slonim and N. Friedman and N. Tishby, Unsupervised document classification using Sequential Information Maximization, in *Proc. of the 25th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval SIGIR*, 2002

3. G. Chechik. and N. Tishby, Extracting relevant structures with side information, *Advances in Neural Information Proceedings Systems NIPS*, 2002
4. 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
5. 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
6. 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
7. 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
8. 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 multi-modal approach for story segmentation of large news video corpus, *TrecVid Workshop)*, 2003
9. J.L. Gauvain, L. Lamel, and G. Adda., The LIMSI Broadcast News Transcription System., *Speech Communication*, 37(1-2):89-108, 2002
10. 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
11. R. Shapire and Y. Singer, Improved boosting algorithms using confidence-rated predictions, in *Proceeding of the Eleventh Annual Conference on Computational Learning Theory*, 1998
12. A. Torralba and K. P. Murphy and W. T. Freeman, Contextual models for object detection using boosted random fields, *NIPS*,2004