TOWARDS SEMANTICALLY MEANINGFUL FEATURE SPACES FOR THE CHARACTERIZATION OF VIDEO CONTENT

Nuno Vasconcelos *Andrew Lippman*

MIT Media Laboratory
{nuno,lip}@media.mit.edu

ABSTRACT

Efficient procedures for browsing, filtering, sorting or retrieving pictorial content require accurate content characterization. Of particular interest are representations based on semantically meaningful feature spaces, capable of capturing properties such as violence, sex or profanity. In this work we report on a first step towards this goal, the design of a stochastic model for video editing which provides a transformation from the image space to a low-dimensional feature space where categorization by degree of action can be easily accomplished.

1. INTRODUCTION

Given the recent growth of information and entertainment delivery services based on direct broadcasting satellites and the future prospects for video delivery by data networks such as the Internet, it is not risky to predict that the living-room of tomorrow will provide access to a multitude of channels carrying diversified video content. In such a setting, the simple task of finding the programs that best suit one's own interests will become daunting, requiring assistance from smart automated information and entertainment appliances.

Such appliances can, however, only become a reality if powerful methods are developed for content characterization, browsing, filtering, and sorting. While a significant effort is being currently developed in these areas for purposes of content-based retrieval [4], most approaches rely on image descriptions which are either specific to some problem domains (such as retrieval of news [9] or faces [6]) or of very low level (such as image color and texture [2]). This limits their scope or/and their capability to support meaningful interaction with users which are not experts in the inner workings of the retrieval systems. In fact, most of the current paradigms for retrieval, such as query by pictorial example or a user-provided scene sketch [3], are far from the semantic representations which most people use to categorize content.

We are interested in the design of feature spaces which can capture high-level scene properties, such as the degree of action, presence/absence of people, or indoors/outdoors set, and semantics such as degree of violence, sex or profanity. In this paper, we report on a first step towards this goal, the design of a statistical model for video editing which provides a transformation from the image space to a low dimensional feature space where categorization by degree of action can be easily accomplished. The premises are simple - action movies have a strong component of short shots with significant inter-frame variation while other types of content typically consist of longer shots with less activity - leading to a feature space characterized by average shot activity and duration. These features are simple to compute (a vital property when dealing with the thousands of frames which compose a typical movie) and seem to provide surprisingly good discrimination. Even more surprisingly, the population of the space by the movies in our database seems to be a good indicator for the degree of violence of their content, leading to a continuous violence scale which is more satisfactory than the binary scales of existing rating systems.

2. MEASURING SHOT PROPERTIES

Our feature set consists of two shot characteristics: average activity and duration. Given a video stream, the computation of these properties is performed in the three steps depicted by Fig. 1.

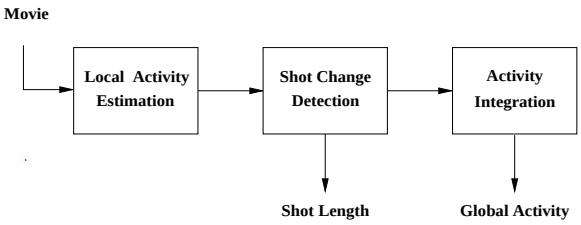


Figure 1: Steps performed for the computation the activity/duration features.

An estimate of local scene activity is first computed for each pair of frames in the sequence. The resulting activity estimates are then passed through a shot boundary detector, responsible for the segmentation of the video stream into its component shots. Next, the activity measures are integrated for each shot and, finally, the shot activities are averaged into an overall measure of the sequence activity. This and the average shot length, computed during shot segmentation, characterize the video stream.

2.1. Estimation of local activity

Given two images of a scene, a considerable number of methods can be used to obtain an estimate of the amount of

variation in the scene. These include simple image subtraction, the energy of the residual error after motion compensation, or the distance between image histograms. Because we are interested in determining how much of the difference between successive frames is due to action in the scene (as opposed to variation due to camera motion or changes of lighting) we rely on the *tangent distance* [5] between the images.

The key idea behind the tangent distance is that, when subject to spatial transformations, images describe manifolds in a very high dimensional space, and a metric invariant to those transformations should measure the distance between those manifolds instead of the distance between other properties of (or features extracted from) the images themselves. However, because the manifolds are very complex, minimizing the distance between them can be a hard optimization problem. The problem can, nevertheless, be made tractable by considering instead the minimization of the distance between the tangents to the manifolds.

Given two images $M(\mathbf{x})$ and $N(\mathbf{x})$, and a transformation $T_{\mathbf{q}}$ parameterized by the vector $\mathbf{q}$, the distance between the associated manifolds is

$$\mathcal{D}(M,N) = \min_{\mathbf{p},\mathbf{q}} ||T_{\mathbf{q}}[M(\mathbf{x})] - T_{\mathbf{p}}[N(\mathbf{x})]||^2. \quad (1)$$

Assuming, for simplicity, that one of the images (M) is fixed, and replacing $T_{\mathbf{p}}[N(\mathbf{x})]$ by the tangent hyper-plane at the point $N(\mathbf{x})$ we obtain the (one-sided) tangent distance

$$\mathcal{D}(M,N) = \min_{\mathbf{p}} ||M(\mathbf{x}) - N(\mathbf{x}) - (\mathbf{p}-\mathbf{I})^T \nabla_{\mathbf{p}} T_{\mathbf{p}}[N(\mathbf{x})]||^2. \quad (2)$$

Many transformations can be used in this equation. Because we are mostly interested in invariance against activity due to camera motion, we consider the set of affine transformations $T_{\mathbf{p}}[N(\mathbf{x})] = N(\psi(\mathbf{x},\mathbf{p}))$, with

$$\psi(\mathbf{x},\mathbf{p}) = \begin{bmatrix} x & y & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & 1 \end{bmatrix} \mathbf{p} = \mathbf{\Phi}(\mathbf{x})\mathbf{p}, \quad (3)$$

capable of compensating for translation (panning), scaling (zooming), in-plane rotation, and shearing. The cost function of equation (2) can be minimized using a multiresolution variant of Newton's method, leading to the following algorithm [7]. For a given level l of the multiresolution decomposition:

1. Compute $N'(\mathbf{x})$ by warping the pattern to classify $N(\mathbf{x})$ according to the best current estimate of $\mathbf{p}$, and compute its spatial gradient $\nabla_{\mathbf{x}} N'(\mathbf{x})$.
2. Update the estimate of $\mathbf{p}_l$ according to

$$\mathbf{p}_l^{n+1} = \mathbf{p}_l^n + \alpha \left[\sum_{\mathbf{x}} \mathbf{\Phi}(\mathbf{x})^T \nabla_{\mathbf{x}} N'(\mathbf{x}) \nabla_{\mathbf{x}}^T N'(\mathbf{x}) \mathbf{\Phi}(\mathbf{x})^T\right]^{-1} \times \left[\sum_{\mathbf{x}} \left[M(\mathbf{x}) - N'(\mathbf{x})\right] \mathbf{\Phi}(\mathbf{x})^T \nabla_{\mathbf{x}} N'(\mathbf{x})\right]. \quad (4)$$

3. Stop if convergence, otherwise go to 1.

Once the final $\mathbf{p}_l$ is obtained, it is passed to the multiresolution level below, by simply doubling the translation parameters. The rescaled vector is then used as initial estimate at the level $l+1$, and the process above repeated.

Once this iterative procedure has converged for all levels of the multiresolution decomposition, the tangent distance between the images is computed through equation (2), using the optimal parameter vector $\mathbf{p}$.

2.2. Shot segmentation and activity integration

The algorithm of the previous section is used for the computation of the local activity measures $\mathcal{D}_f$, $f = 1,\ldots,F-1$, where F is the number of frames in the video stream. Shot segmentation is then performed according to a Bayesian model of the editing process which incorporates prior knowledge about the shot duration [8]. Under this model a shot boundary is declared whenever

$$\log \frac{P(\mathcal{D}_f|\mathcal{S}_f=1)}{P(\mathcal{D}_f|\mathcal{S}_f=0)} \geq \mathcal{T}(\tau_f), \quad (5)$$

where $\mathcal{S}_f$ is an indicator variable which takes the value 1 whenever a shot boundary is present and 0 otherwise, and $\mathcal{T}(\tau_f)$ is an adaptive threshold which is a function of the prior density for shot duration and the time τ_f elapsed since the previous boundary.

In [8] we show that the *Weibull* density has appealing properties as a model for the shot duration, leading to the threshold

$$\mathcal{T}(\tau_f) = -\log\left\{\exp\left[\frac{(\tau_f+\delta)^\alpha - \tau_f^\alpha}{\beta^\alpha}\right] - 1\right\}, \quad (6)$$

where α, β, and δ are density parameters. This threshold is easy to compute, and has the intuitive behavior illustrated by figure 2: while in the initial segment of the shot $T(\tau_f)$ is very high and shot changes are very unlikely to be accepted, $T(\tau_f)$ decreases as the scene progresses increasing the likelihood that shot boundaries will be accepted.

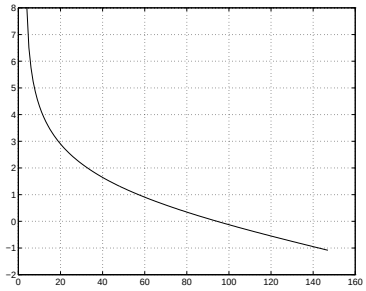


Figure 2: Shot detection threshold for a Weibull prior.

Given the shot segmentation, the activity of each shot is computed by $\mathcal{S}_s = f(\mathcal{D}_i), i \in s$, where s is the shot number. Currently, we use average for f, but other functions (including maximum, minimum, and median) could be used as well.

Finally, the video stream is characterized by the feature pair $(\mathcal{A}, \mathcal{L})$, where $\mathcal{A}$ and $\mathcal{L}$ are, respectively, the average shot activity and length.

3. ANALYSIS OF EXPERIMENTAL DATA

To test the validity of the transformation into the above feature space as a means of discriminating among diverse types of video content, we applied it to a database of 23 promotional movie trailers. Each trailer is a summary of the corresponding movie with average length of approximately two minutes, and the entire database requires around 26 Giga bytes of storage. The names of the movies are listed on the table of Fig. 3.

3.1. Action characterization

The figure shows how the movies populate the feature space. A search in the *Internet Movie Database* (IMDB) [1] revealed that none of the movies above the lower line depicted in the figure include *action* as one of their descriptive genre keywords. On the other hand, of those below the line, only the comedy "Blankman" and the thriller "Madness" did not include the *action* keyword, even though they could have been easily classified as *action/comedy* and *action/thriller*, respectively.

The upper line in the figure also appears to provide a separation of the space which is semantically meaningful. While all the movies above this line contained either *comedy*, or *romance*, or both as genre keywords, of those below it only "Jungle" was categorized as *romance* and "Edwood" and "Blankman" as *comedies*.

Further investigation revealed the reason for these outliers: while the comedies above the line are typically categorized as *comedy/romance* or simply *comedy*, "Edwood" receives the awkward categorization of *comedy/drama* (indicating that characterizing its content is probably a difficult task), and "Blankman" that of *comedy/screwball/super hero* confirming the fact that it is an action-packed comedy, which could easily fall in the *action* category. On the other hand, while the romances above the line either belong to the category *drama/romance* or *comedy/romance*, "Jungle" is categorized as *adventure/romance* indicating a degree of action which is above that of the other movies in the *romance* class.

It seems, therefore, that this is a feature space where the movies are nicely laid out according to the degree of action in their scripts, providing discrimination between several genres of content. For example, and even though definitive conclusions cannot be taken from such a small database, the dataset of Fig. 3 suggests that even a simple Gaussian classifier would achieve high classification accuracy for the task of detecting action movies.

3.2. Violence characterization

It is also interesting to note that the scene length/activity space seems to be a good feature space for the categorization of movies according to the violence. In addition to genre keywords we also extracted from the IMDB the rating assigned to each movie by the Motion Picture Association of America, as well as the reasons given to support such ratings. Of all the movies in our database, only two contained the word *violence* as a reason for the received rate: "Dredd" and "Vengeance". This would clearly be too coarse of a quantization for a query based on the amount of violence.

On the other hand, the feature space of Fig. 3 provides a nice clustering of the movies according to their violence. In the graph, not only "Vengeance" is clearly singled out from the rest of the pack, but "Street Fighter" (a movie after the video game of the same name), "Madness" and "Terminal" also correctly stand out as violent movies. On the opposite corner of the space, are the romances and comedies, containing non-violent content, and as one progresses towards the violent corner one encounters titles which are increasingly more likely to contain violence.

The categorization is, however, not perfect. "Dredd", for example, is not identified as one of the most violent entries in the database. But, also here, there is a reason for the apparent outlier: perhaps in an effort to tone down the violent nature of the movie, its promotional trailer has a component of action scenes which is much smaller than one would expect. This stresses the point that examination of short portions of a movie (even if summaries such as the trailers that we have analyzed) are not guaranteed to always provide accurate results.

4. DISCUSSION

The main advantage of relying on a graphical layout such as that of Fig. 3 to rate video content according to action or violence is that it provides a continuous scale, as opposed to the binary scales currently available. Instead of being limited to queries of the type "show me a movie which is violent/not-violent", a continuous scale supports queries such as "show me all movies whose violence is between those of movie X and movie Y", providing a much richer content-description than simple categorization into a few classes.

Obviously, a system based on features as simple as the ones considered above cannot be expected to achieve perfect classification for a sophisticated concept such as degree of violence. However, the results above suggest that such a system would be right most of the time, and could be a useful complement to the current rating paradigms.

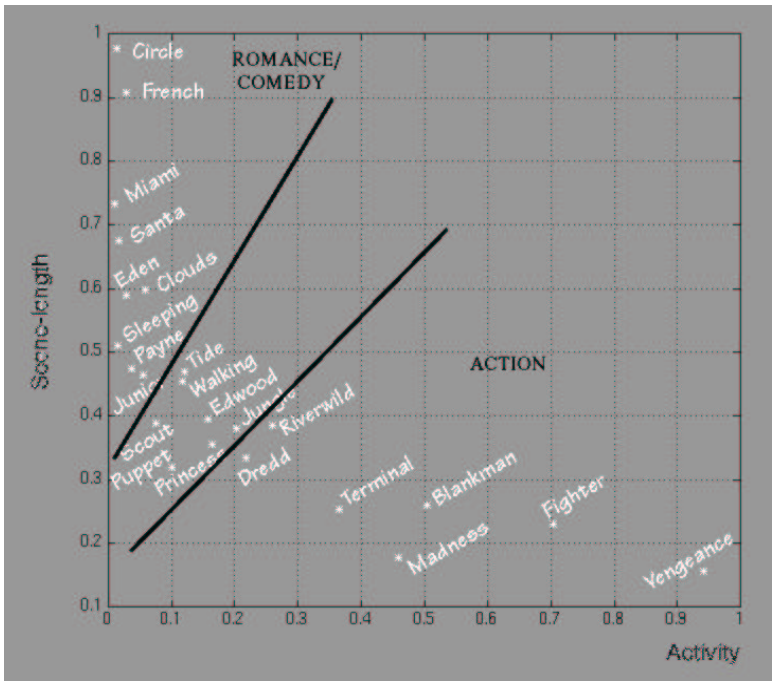
5. REFERENCES

[1] *Internet Movie Database*. http://us.imdb.com/.

[2] Y. Gong, H. Zhang, H. Chuan, and M. Sakauchi. An Image Database System with Content Capturing and Fast Image Indexing Abilities. In *Proc. Int. Conf. on Multimedia Computing and Systems*, May 1994, Boston, USA.

[3] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content using color, texture, and shape. In *SPIE Storage and Retrieval for Image and Video Databases, San Jose, California*, pages 173–181, 1993.

[4] R. Picard. Light-years from Lena: Video and Image



Legend	Movie
Circle	"Circle of Friends"
French	"French Kiss"
Miami	"Miami Rhapsody"
Santa	"The Santa Clause"
Eden	"Exit to Eden"
Clouds	"A Walk in the Clouds"
Sleeping	"While you Were Sleeping"
Payne	"Major Payne"
Junior	"Junior"
Tide	"Crimson Tide"
Scout	"The Scout"
Walking	"The Walking Dead"
Edwood	"Ed Wood"
Jungle	"The Jungle Book"
Puppet	"Puppet Master"
Princess	"A Little Princess"
Dredd	"Judge Dredd"
Riverwild	"The River Wild"
Terminal	"Terminal Velocity"
Blankman	"Blankman"
Madness	"In the Mouth of Madness"
Fighter	"Street Fighter"
Vengeance	"Die Hard: With a Vengeance"

Figure 3: Population of the feature space by the movies in our database. Movie names are listed in the table on the right.

Libraries of the Future. In *Proc. Int. Conf. Image Processing*, October 1995, Washington DC, USA.

[5] P. Simard, Y. Le Cun, and J. Denker. Memory-based Character Recognition Using a Transformation Invariant Metric. In *Int. Conference on Pattern Recognition*, Jerusalem, Israel, 1994.

[6] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3, 1991.

[7] N. Vasconcelos and A. Lippman. Multiresolution Tangent Distance for Affine Invariant Retrieval. Technical report, MIT Media Laboratory, 1997. Available in http://www.media.mit.edu/~nuno.

[8] N. Vasconcelos and A. Lippman. A Bayesian Video Modeling Framework for Shot Segmentation and Content Characterization. In *Proc. IEEE Workshop on Content-based Access to Image and Video Libraries*, CVPR97, San Juan, Puerto Rico, 1997.

[9] H. Zhang, Y. Gong, S. Smoliar, and S. Tan. Automatic Parsing of News Video. In *Proc. Int. Conf. on Multimedia Computing and Systems*, May 1994, Boston, USA.