# Chapter 5

# Conclusions

The high-level semantic structure of video can be automatically parsed using domain-specific knowledge. This knowledge can be expressed in many ways, but always based on semantic concepts. This intermediate-level semantics is usually obtained in terms of the requirements and constraints imposed by the domain. The CMC-based representation of visual contents in video developed in this thesis provides intermediate-level semantics without considering domain knowledge. Semantically meaningful clusterings can be automatically obtained, so that only higher level knowledge must be described and applied in order to obtain the semantic structure of the video.

Common approaches to video structure analysis have been based on representing shots by one or several keyframes and computing low-level and intermediate-level descriptors of their contents. Then, semantically meaningful clusterings are obtained using specifically tailored shot similarity measures based on their descriptions. One main advantage of the CMC approach over other representations of visual contents in video is that shot contents are seen as temporal processes, thus information from all the images of the shot is considered. Also, natural similarity measures are provided by the probabilistic framework.

The CMC approach to contents representation can summarize the contribution of several features, as well as their possible relationships and dependencies. We have presented a method to obtain the optimal coupling structure in terms of minimum cost and minimum loss of information, which has been proven to be directly related to the accuracy of the representation. This method also tells us that an approach to representing the contents of an image sequence based on accumulating static image descriptions is not appropriate. The temporal behavior of low-level features is highly informational.

During this thesis, we have seen that color and motion are two major contributors to obtain intermediate-level semantics. Color provides information about objects, location and even emotional aspects attached to contents. On the other hand, motion provides information about the type of shot in terms of relative distance of the camera to the subject matter, camera operation and more complex concepts like crowds or talking heads. The combination of color and motion information in a CMC based representation of shots provides intermediate-level semantics about objects (identity, size and motion), camera operation, location, type of shot, temporal relationships between elements of the scene, and global activity (understood as the "amount of action"). More complex semantic concepts emerge from the combination of this intermediate-level information, like the already mentioned "crowds" and "talking heads", which can be sub-classified into "anchor shots" and "correspondents and interviewees" in the domain of News. This information is not only useful for high-level structure analysis, but also for automatic or assisted video indexing and annotation.

The representation capabilities of the CMC approach have been shown for the Sports (soccer) and News domains. In the latter, its high-level structure has been obtained by defining very simple rules about the domain. During the process of computing shot representations, shot boundaries are detected as well. Therefore, the same CMC representation integrates all the steps of the process up to the intermediate-level semantic clustering of shots.

## 5.1 Other contributions

The main contribution of this thesis is the CMC-based modelig for representing visual contents in video sequences, and its application to news structure extraction, shot boundary detection, object localization and video retrieval. Other contributions also presented in this dissertation are the following:

**CECA algorithm:** The CECA algorithm for shot boundary detection combines color and edges information to obtain a more robust shot detector than usual approaches, in terms of high precision and recall for both abrupt and gradual transitions.

**AudiCom:** AudiCom is a system for the automatic recognition of video segments in a longer stream, which has been applied to TV commercials monitoring. It is an example of keyframe-based representation of shot contents.

**GMM-based color correction:** Mappings between different device-dependent color appearances are defined in terms of the parameters of Gaussian distributions. This method implicitly defines an intermediate normalized color appearance space.

**Semiotic analysis:** Semiotics provides semantic information about the use of low-level features to convey emotional aspects of video contents and production. Color, motion, orientations and other features are attached to semantic concepts, which can be exploited from the machine vision standpoint.

**Semantics from motion:** Color has been considered one of the main semantic carriers. The semantics conveyed by low-level motion information have been analyzed in order to evaluate what kind of information can be obtained from it, and its usefulness in terms of intermediate-level semantics.

## 5.2 Future work

The work developed in this thesis can be continued and extended in different ways. First of all, through the entire thesis we have considered that the set of cliques of the 3D MRF associated to the pixels of an image sequence is formed by every pair of sites with the same spatial position and consecutive time instants. However, we could also consider the motion of the pixels. The pixel located at $(x, y)$ in image $I_t$ may be found at a different position $(x', y')$ in image $I_{t+1}$ due to global or local motion. In this case, the transition from the states at $I_t(x, y) \rightarrow I_{t+1}(x', y')$ should be considered instead of $I_t(x, y) \rightarrow I_{t+1}(x, y)$. In this way, we would not only obtain a summary of the temporal behavior of low-level features, but we would also keep track of their actual changes along time. However, this would require an accurate estimation of a dense optical flow field, which is costly and difficult to obtain.

Global motion can be sometimes very useful, as it provides information about camera operation and activity. However, the motion generated by the actions performed by objects and other important foreground elements can be blurred or misleading due to the presence of global motion. Global motion compensation prior to the computation of CMC model parameters will eliminate camera operation information from the representation, but may let us obtain more accurate representations of actions in terms of local motions. In this way, the action classification problem could be faced.

Color and motion features have been used in this work, as they are known to convey very meaningful semantic information. Other features like texture and shape should also be analyzed in order to identify possible relationships to intermediate-level semantic concepts. In terms of performance, MPEG compression standards are based on the computation of DCT coefficients, which are related to luminance and chroma, and motion vectors. These already computed features could be used to quickly obtain the CMC representation of shots and directly analyze MPEG-compressed videos. Also, the combination of more than 2 features might provide more robust and informative representations of shot contents.

Simple rules defined using domain knowledge can take us from semantically described shots to the structural “story unit” level. A story unit is a scene, a play or a news item, depending on the domain. There are other levels on top of story units in the hierarchical structure of videos. They can be called news sections (local, international, sports, ...) or game periods. To reach this level of abstraction, more complex knowledge is required. This knowledge may also take advantage of intermediate-level semantics, thus adding one more term to the first equation of this dissertation:

$$\textit{Video structure} = \textit{Domain knowledge} + \textit{Story units} + \textit{Interm.-level semantics}.$$

# Appendix A

# Video color correction using Gaussian mixture models

This appendix gives details about the characterization of device color spaces and the definition of mappings between different device colors using Gaussian mixture models.

## A.1 Color appearance modeling and mapping

Considering the case of single Gaussian distributions, there is an affine transformation that can be expressed in terms of their means $\mu_1$, $\mu_2$ and the spectral decomposition of their covariance matrices $\Sigma_1$, $\Sigma_2$. The eigenvalues $\lambda_{ij}$ of $\Sigma_i$ are the variances of each distribution along their principal axes, which are given by their corresponding eigenvectors $e_{ij}$. Using these parameters, the final transformation matrix $\mathbf{A}$ is given by a composition of translation ($\mathbf{T}$), rotation ($\mathbf{R}$) and scaling ($\mathbf{S}$) matrices:

$$\mathbf{A} = \mathbf{T}_2^{-1}\mathbf{R}_2^{-1}\mathbf{S}_2^{-1}\mathbf{S}_1\mathbf{R}_1\mathbf{T}_1 \tag{A.1}$$

where $\mathbf{T}_i$, $\mathbf{R}_i$ and $\mathbf{S}_i$ are expressed in terms of $\mu_i$, $e_{ij}$ and $\lambda_{ij}$ respectively. Figure A.1 shows the transformation process. Note that the intermediate Gaussian $G(0, \mathbf{I})$ plays the role of device-independent color space. Therefore, we only need to store the transformation matrix $\mathbf{A}_D = \mathbf{S}_D\mathbf{R}_D\mathbf{T}_D$ from device $D$ to the intermediate space, and the direct color transformation between two devices $D_1$ and $D_2$ would be given by $\mathbf{A} = \mathbf{A}_{D_2}^{-1}\mathbf{A}_{D_1}$.

The main problem is finding the right assignment between the axes of both distributions. We assume that the ratios between the variances of the different axes are kept under different sensors. Therefore, the principal eigenvectors of the distributions have the same ordering. On the other hand, each principal axis can be expressed either by an eigenvector or by its opposite direction. This could lead to a 180-degree rotation

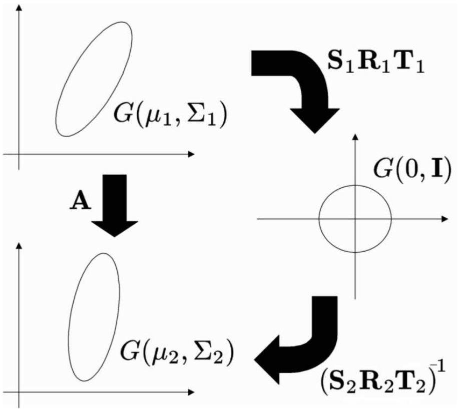

$G(\mu_1, \Sigma_1)$
$\mathbf{S}_1\mathbf{R}_1\mathbf{T}_1$
$G(0, \mathbf{I})$
$\mathbf{A}$
$G(\mu_2, \Sigma_2)$
$(\mathbf{S}_2\mathbf{R}_2\mathbf{T}_2)^{-1}$

**Figure A.1:** Transformation of a data set in order to obtain a different Gaussian distribution.

on one of the axis of the distribution and colors would not be mapped correctly. As long as we know the correspondence of axes between both distributions, we can avoid these situations by heading corresponding eigenvectors for the same direction.

The extension to the case $\{G(\mu_i^{(1)}, \Sigma_i^{(1)}), ..., G(\mu_i^{(n)}, \Sigma_i^{(n)})\}$ of $n$ Gaussians in each mixture, where $G(\mu_i^{(j)}, \Sigma_i^{(j)})$ denotes the $j$th Gaussian component of the model for device $i$, requires that:

- a unique correspondence between the Gaussian components of both mixtures exists,
- and spatial relationships between mixture components in RGB space are kept in all distributions.

The full transformation is then expressed by the set of matrices $\{\mathbf{A}_1, ..., \mathbf{A}_n\}$ that transform each component of the mixture. Given a RGB color $x = (r, g, b, 1)^T$ to be mapped, the transformation matrix corresponding to the maximum probability component of the mixture should be applied:

$$x' = \mathbf{A}_j x, \quad j = \arg \max_{k=1..n} P(G(\mu_1^{(k)}, \Sigma_1^{(k)})|x) \tag{A.2}$$

where $P(G(\mu_1^{(k)}, \Sigma_1^{(k)})|x)$ is the posterior probability of the $k$th mixture component, given the sample $x$, computed using Bayes rule as:

$$P(G(\mu_1^{(k)}, \Sigma_1^{(k)})|x) = \frac{P(x|G(\mu_1^{(k)}, \Sigma_1^{(k)}))P(G(\mu_1^{(k)}, \Sigma_1^{(k)}))}{\sum_{k=1}^{n} P(x|G(\mu_1^{(k)}, \Sigma_1^{(k)}))P(G(\mu_1^{(k)}, \Sigma_1^{(k)}))} \tag{A.3}$$

Assuming the same prior probability for each Gaussian component, $j$ is also obtained by maximizing their likelihood $P(x|G(\mu_1^{(k)}, \Sigma_1^{(k)}))$.

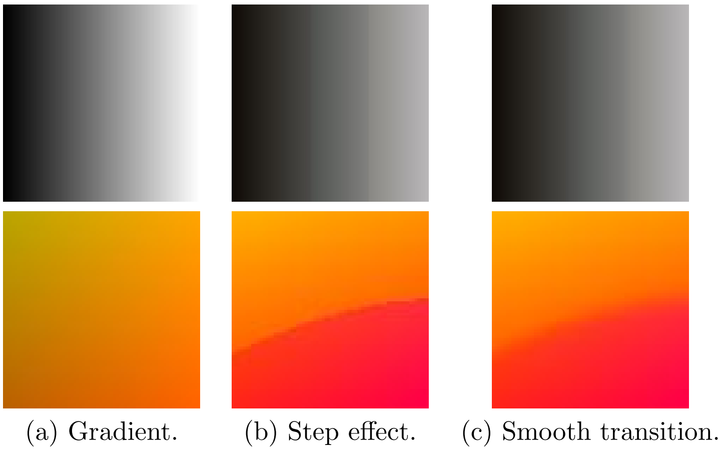

(a) Gradient. (b) Step effect. (c) Smooth transition.

**Figure A.2:** (a) Grayscale (top) and color (bottom) gradients. (b) Step effects appear when they are mapped using MAP classification in the transformation. (c) Smooth transitions are obtained when different transformations are combined in terms of their probabilities.

However, a significant step effect may appear in edges and color gradients due to the loss of continuity produced by the *maximum a posteriori* (MAP) classifier in eq. (A.2), which may assign different transformations to nearby colors. This effect is shown in grayscale and color gradients in fig. A.2(b). In order to obtain a smooth transition between the different mappings applied in various regions of the color space, the contribution of the different parts of the transformation should be combined taking into account their probability of being applied:

$$x' = \sum_{j=1}^{n} P(G(\mu_1^{(j)}, \Sigma_1^{(j)})|x)\mathbf{A}_j x \tag{A.4}$$

Figure A.2(c) shows the smoothness of the transition between different transformations that are applied to the various regions of the color space. The mapping defined by eq. (A.4) is intended to preserve the intrinsic appearance of colors. Each color concept contributes to the transformation with its representativity of the particular color that is being considered.

The final transformation requires the computation of $n$ 3-dimensional Gaussian probabilities and $n$ $4 \times 4$ matrix products. The implementation as a look-up table or using simple specific hardware can be considered in order to allow real time processing in current home multimedia systems.

## A.2 Examples of device color mappings

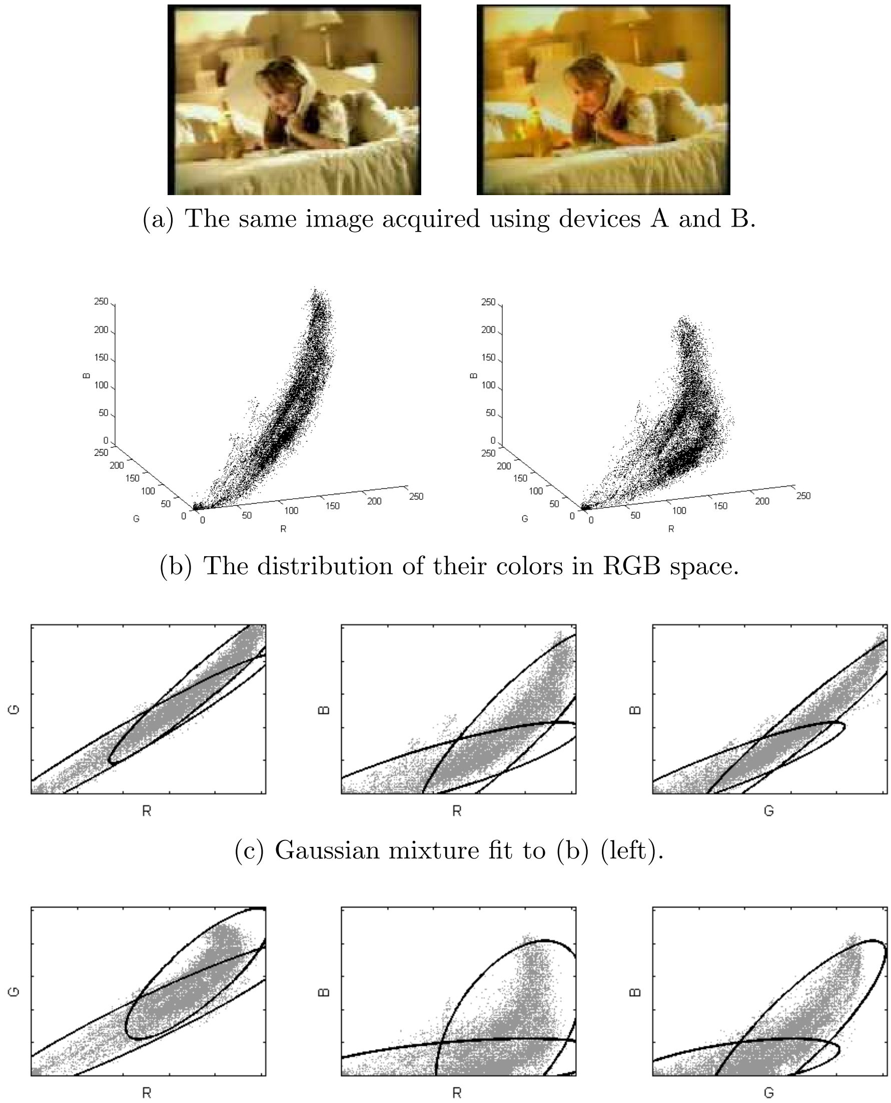
As an example of the proposed color mapping strategy, the simple case of mapping the color appearance of an image acquired using two different combinations of video devices, which we shall call A and B, is considered. The variation of the image gamut given by different devices can be visually assessed in fig. A.3(a), showing a noticeable non-linear distortion along the RGB color space, as seen in fig. A.3(b). Each distribution of colors was parameterized by a mixture of two Gaussians (figs. A.3(c) and A.3(d)), whose parameters were estimated from the data sets using the EM algorithm [15]. In this case, the correspondence between the components of both mixtures is straightforward. Note that the spatial relationships between the Gaussians are kept in both distributions as well, so that all requirements needed to obtain the transformation are fulfilled.

The transformation matrices that map the color appearance of the image acquired using B as if it had been acquired using A were obtained and applied as defined in eq. (A.4). The results are shown in fig. A.4, where it can be seen that the corrected image from device B has acquired the color appearance of device A. The transformation does not give exact results due to other device imperfections, like signal noise and image misalignment. However, the color appearance obtained is perceptually the same.

The extension of the GMM color mapping strategy to the general device gamut case requires modeling the distribution of all possible colors generated in RGB space that can be obtained using a particular combination of video devices. This distribution can be automatically learned from a pattern like the ones shown in fig. A.5, which encodes the means and covariance matrices of a predefined mixture of Gaussians specifically designed to fill up the whole color space. Both requirements of the multiple Gaussians transformation are implicitly fulfilled by computing the parameters of each component from its corresponding region in the pattern. In this way, a parametric model of the color distribution generated by a particular device is obtained by acquiring this pattern.

The number and location of the mixture components are not imposed by the method, whenever the joint distribution considers all possible colors that can be generated by a device. The non-linearities of the transformations are better captured using as many Gaussians as possible, which also increases the computational cost of the mapping. Different configurations may suit different color imaging applications. However, we are trying to consider the most general-purpose case. Note that the number and location of these basic colors are chosen following a spatial criterion, in order to cover as much as possible from the RGB space. Commonly used color calibration patterns like the Macbeth ColorChecker provide an arbitrary number of color patches chosen by their spectral nature instead [48], which is not our goal as we are not characterizing the spectral responses of sensors. We show examples using different configurations:

1. 1 component in the middle of the RGB space (model for gray) and 8 equally spaced components in the periphery (models for black, white, red, green, blue,



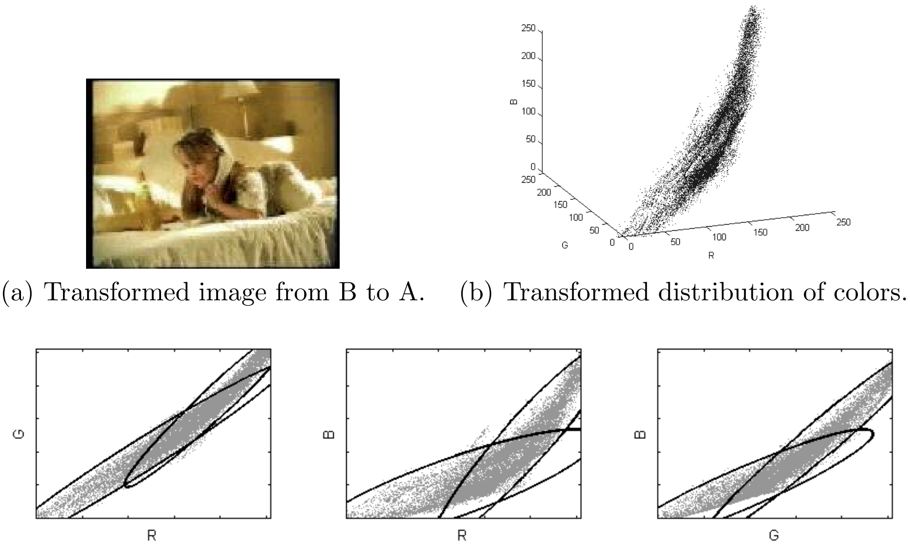
(a) The same image acquired using devices A and B.

(b) The distribution of their colors in RGB space.

(c) Gaussian mixture fit to (b) (left).

(d) Gaussian mixture fit to (b) (right).

**Figure A.3:** The same image acquired using different video devices shows a significant non-linear variation of its color distribution.



(a) Transformed image from B to A. (b) Transformed distribution of colors.

(c) Gaussian mixture fit to (b).

**Figure A.4:** Result of applying the transformation between the distributions in fig. A.3.

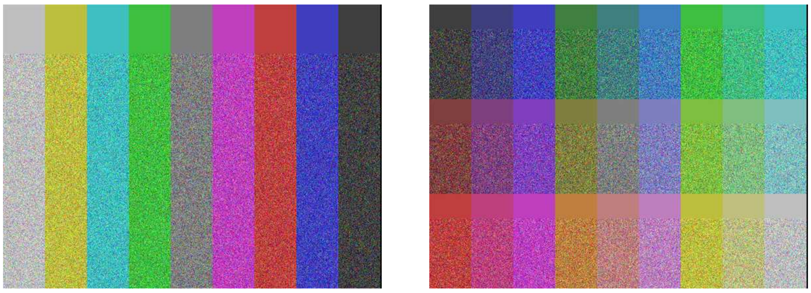
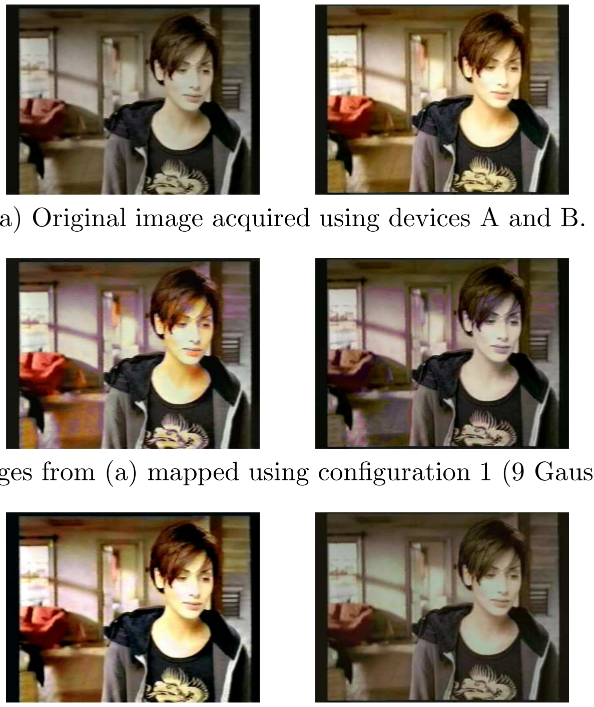


**Figure A.5:** Patterns encoding the parameters of the mixture of 9 (left) and 27 (right) Gaussians that will encode device-dependent color appearances. Flat regions encode Gaussian means, and noisy regions encode their covariance matrices.



(a) Original image acquired using devices A and B.

(b) Images from (a) mapped using configuration 1 (9 Gaussians).

(c) Images from (a) mapped using configuration 2 (27 Gaussians).

**Figure A.6:** Image color appearance transformations using global models with different configurations.

cyan, yellow and magenta), encoded by the pattern in fig. A.5(left),

2. a 3 × 3 × 3 grid of equally spaced components, encoded by the pattern in fig. A.5(right).

Corrections using the first one showed some unnatural colors, which can be noticed in fig. A.6(b). On the other hand, the higher representation capability of the second configuration gave much better results, as can be seen in fig. A.6(c). This way of defining device gamut mappings allows us to obtain color appearances as given by different devices without prior knowledge about its final distribution in RGB space.

## A.3 Comparison with other color correction methods

The GMM approach has been compared to known color normalization algorithms. The comparison was performed in the context of AudiCom. This system uses a keyframe representation of video shots, which are compared to those in a database of TV commercials. Therefore, the problem is posed as a matching of still images. Principal Component Analysis (PCA) of the keyframes was used to obtain a low-dimensional representation space, where matching was performed by minimum Euclidean distance.

Many color normalization methods that allow color image independence from different image formation factors have been developed in the computer vision literature. Amongst these factors, the one that seems to be more suitable to the variation produced by the change of video acquisition device is the color variation of the illuminant. In this sense, we have considered color constancy algorithms that allow independence from the illuminant color following a diagonal model. Diagonal models consist of a single scale factor per color channel, and have been shown to work almost as well as full 3 × 3 linear models [20].

The grayworld algorithm takes its name from the assumption that the average color in all images is an ideal or canonical gray. The scale factors are $R_g/\bar{R}$, $G_g/\bar{G}$, and $B_g/\bar{B}$, where $(R_g, G_g, B_g)$ is the canonical gray $RGB$ value and $(\bar{R}, \bar{G}, \bar{B})$ is the average color of the image. Following this approach, the white-patch retinex algorithm scales the colors of the image with respect to the canonical white. The scale factors in this case are $R_w/R_{max}$, $G_w/G_{max}$, and $B_w/B_{max}$, where $(R_w, G_w, B_w)$ is the canonical white and $(R_{max}, G_{max}, B_{max})$ are the largest values in each color channel of the image.

The comprehensive color image normalization (CCN) by Finlayson et al. [22] has also been considered in this work as a different approach to color normalization. This algorithm removes image dependency on lighting geometry and illuminant color, by applying functions $R$ and $C$ defined as:

$$R(I)_{i,j} = \frac{I_{i,j}}{\sum_{k=1}^{3} I_{i,k}} \tag{A.5}$$

$$C(I)_{i,j} = \frac{N/3 I_{i,j}}{\sum_{k=1}^{N} I_{k,j}} \tag{A.6}$$

where $I$ is an $N \times 3$ image matrix, whose columns contain the intensity of the image pixels in the 3 $RGB$ color channels. The normalization procedure is defined as an iterative process until no change is detected in the image:

1. $I_0 = I$
2. Do $I_{i+1} = C(R(I_i))$ until $I_{i+1} = I_i$

Note that $R$ removes image dependency on the intensity of the illuminant and $C$ is equivalent to the grayworld correction, thus giving an illuminant color independent image.

The database used was composed of 220 keyframes acquired from two different music videos using a professional JVC BR-S822E VCR and a miroVideo DC30 Plus

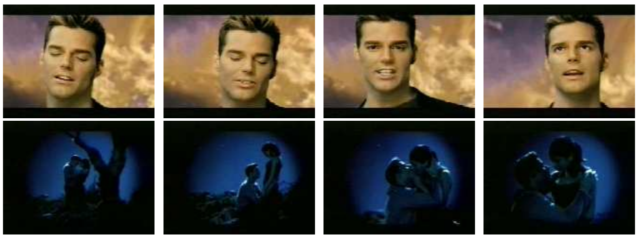


**Figure A.7:** Hardly discriminable images in the database.

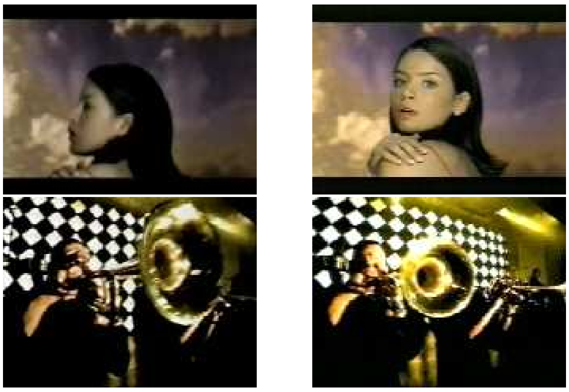


**Figure A.8:** Non-exact matches between test (left) and database (right) images.

video digitizer card. The keyframes were automatically extracted from the video stream using the shot segmentation capabilities of AudiCom. Due to the way music videos are edited, some sets of hardly discriminable keyframes, which actually correspond to different shots, are found in the database. Two of these sets are shown in fig. A.7.

Test images were acquired from the same video-clips using a home Mitsubishi M1000 VCR and a Matrox Marvel G200 TV video acquisition card. It is interesting to note that the automatic keyframe selection process may give a slightly different number of keyframes, and even in different positions, depending on CPU and other resources usage, which lead to variations between test and database images like the ones shown in fig. A.8. We decided to keep these non-exact matches because they may happen in the real application as well. The final test set was composed of 183 images.

The results of applying appearance based recognition of the test keyframes with respect to the ones in the database using the color normalization schemes considered in this work are shown in fig. A.9. Nothing denotes doing no normalization at all, while GMM stands for Gaussian mixture models color normalization.

The rate obtained using GMM color normalization is always higher than with no normalization process. It is important to note that the highest recognition rate

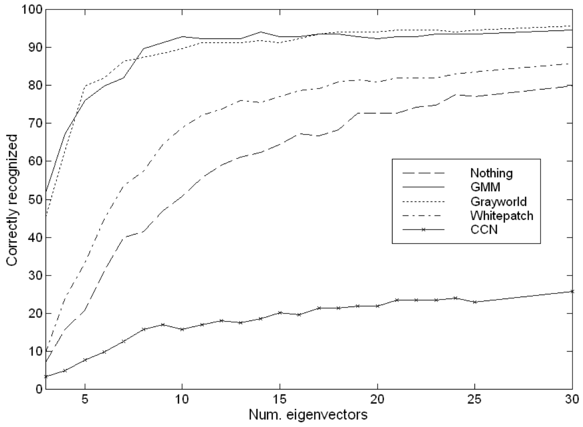


**Figure A.9:** Appearance based recognition rates as a function of the number of dimensions of the representation space, after applying different color normalization algorithms.

| | Original | GMM normalized |
|---|---|---|
| Hist. Inters. | 6.56% | 40.98% |
| PCA 3 eigens | 7.1% | 51.91% |
| PCA 10 eigens | 50.81% | 92.89% |
| PCA 30 eigens | 79.78% | 94.53% |
| PCA 50 eigens | 81.96% | 93.44% |

**Table A.1:** Recognition rates obtained using histogram intersection and different configurations of appearance based coding.

obtained without normalization was 81.9%, which is hardly acceptable in common recognition applications. On the other hand, the maximum rate obtained using GMM normalization reached a fairly good 94.5%. Moreover, the use of this color normalization allows rates close to the maximum to be reached keeping 10 eigenvectors, which means preserving only 53.27% of the total variance of the original image set. Therefore, our color normalization makes the global color appearance of test images to be much more similar to their corresponding ones in the database than if no normalization is applied.

These results are also compared in table A.1 to the ones obtained using color histogram intersection. The nature of the images in the database (see fig. A.7) let us anticipate bad results using this method. In fact, only 3 components in the appearance based representation are needed to outperform it. However, there is an impressive increase in the recognition rate when the GMM color normalization is applied.

Some result images obtained using the 30 components PCA representation are also shown in fig. A.10. The most interesting cases of hardly discriminable keyframes

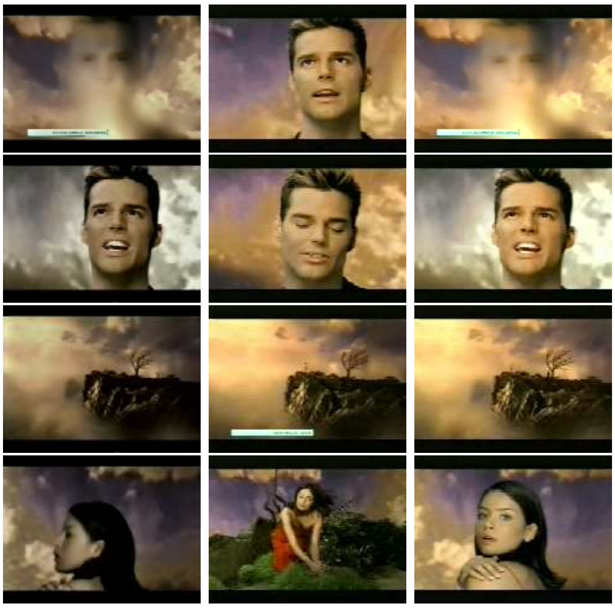
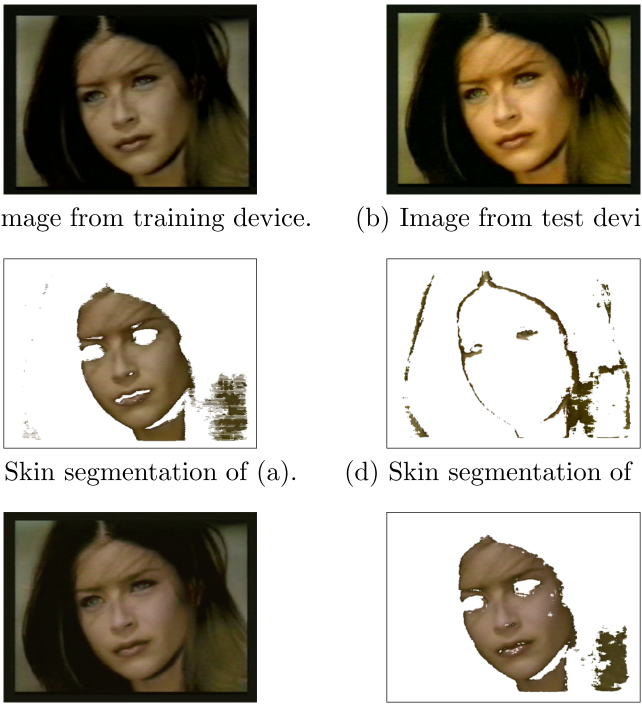


**Figure A.10:** Recognition results. Keyframes to be recognized (left), and the results without (middle) and with (right) GMM color normalization.

and non-exact matching are given in the third and four rows, respectively.

On the other hand, comparing the performance obtained using the GMM normalization and the other approaches considered, results show that the grayworld color constancy algorithm reaches almost identical recognition rates as the GMM approach. This algorithm seems to be very well suited to the color variation caused by the change of device in our experiments. The whitepatch retinex approach slightly improves the recognition performance with respect to doing no normalization at all, but the rates obtained (less than 90%) are not enough for most common applications. The poor performance provided by the CCN algorithm is quite surprising, considering that one of its steps is equivalent to the grayworld algorithm. These results show that not all color normalization algorithms that work well in the presence of illuminant color variations can give as good results when color variation is caused by the change of video acquisition device.

We want to stress the idea that the GMM approach preserves the underlying colors of the scene contents, as it is a color appearance-based normalization. To show this, we have applied the method to skin color segmentation. In this case, skin color samples for training are taken from one particular device, while test images are acquired using a second device. Results can be seen in fig. A.11. The concept of skin color is preserved under the GMM-based transformations, while it may vary from one image to the others when the grayworld, whitepatch or CCN algorithms are used. These algorithms fully depend on the specific image that is being normalized, while the GMM approach is defined by device characteristics. The normalization applied



(a) Image from training device. (b) Image from test device.

(c) Skin segmentation of (a). (d) Skin segmentation of (b).

(e) Color adapted image (b). (f) Skin segmentation of (e).

**Figure A.11:** Example application of the GMM color mapping to skin color segmentation.

in this way is the same for every image acquired using a particular device. This fact makes the GMM approach to be more suitable to be used in a general digital video libraries framework.

## A.4 Summary

The GMM approach to the characterization and mapping of device-dependent color spaces has the advantages:

- The characterization of device-dependent color spaces is parametric and model parameters are easily estimated using a particular calibration pattern.
- Color appearance mappings are defined in order to keep the intrinsic appearance of colors, i.e. their identity. A perceptual color space is not required, so that the mappings can be defined in the RGB space.
- The final mapping obtained is non-linear. Different mappings are applied to different regions of the color space, with smooth transitions between them.

- Luminance and chroma components do not need to be processed separately. Axes of the color space are not assumed to be independent.
- Only one mapping per device is required, using the concept of an intermediate normalized color appearance space.

Experiments show that the GMM approach is suitable for appearance-based image retrieval. Unlike other color normalization methods, the GMM approach keeps the underlying identity of colors, as shown in the skin color segmentation application.

# Appendix B

## ML parameter estimation for a fully observed MC

This appendix shows how to estimate the maximum likelihood (ML) parameters of a fully observed Markov chain, i.e. its transition probabilities. Maximizing the log-likelihood, the likelihood is maximized as well. From eq. (3.6), the log-likelihood of a MC is:

$$\mathcal{L}\{P(x)\} = \log P(x_1) + \sum_{(i,j)\in S^2} C_{ij|x} \log T_{ij} \tag{B.1}$$

Taking the derivatives of the log-likelihood with respect to $T_{ij}$:

$$\frac{\partial \mathcal{L}\{P(x)\}}{\partial T_{ij}} = C_{ij|x} \frac{1}{T_{ij}} \tag{B.2}$$

The final $T_{ij}$'s must fulfill that $\sum_{j\in S} T_{ij} = 1$, as it is a conditional probability distribution. This constraint can be forced adding Lagrangian multipliers:

$$\lambda \left( \sum_{j\in S} T_{ij} - 1 \right) = 0 \tag{B.3}$$

Taking derivatives with respect to $T_{ij}$:

$$\frac{\partial}{\partial T_{ij}} = \lambda \tag{B.4}$$

Combining eqs. (B.2) and (B.4), and equaling to 0:

$$\frac{C_{ij|x}}{T_{ij}} + \lambda = 0 \tag{B.5}$$

$$T_{ij} = \frac{C_{ij|x}}{\lambda} \tag{B.6}$$

Since $\lambda$ must normalize the conditional probabilities, the final ML parameters are given by:

$$\hat{T}_{ij} = \frac{C_{ij|x}}{\sum_{j \in S} C_{ij|x}} \tag{B.7}$$

which is a normalization of temporal cooccurrences.

# Appendix C

# Optimal transition probabilities for the CMC structure

This appendix shows how to compute the optimal transition probabilities for the CMC structure with respect to the Cartesian product (CP) structure in terms of minimum mutual information. The mutual information of the CMC with respect to the CP structure is:

$$D(P||P') = \sum_{i,j,k,l} P(i,j,k,l) \log \frac{P(i,j|k,l)}{P(i|k,l)P(j|k,l)} \tag{C.1}$$

where $P$ and $P'$ are the probability distributions associated to the CMC and the CP structures, respectively. Taking the derivatives with respect to each $P(i|k,l)$ (and $P(j|k,l)$):

$$\begin{aligned}
\frac{\partial D(P||P')}{\partial P(i|k,l)} &= \frac{\partial}{\partial P(i|k,l)} \sum_{i,j,k,l} P(i,j,k,l) \log \frac{P(i,j|k,l)}{P(i|k,l)P(j|k,l)} \\
&= \frac{\partial}{\partial P(i|k,l)} \sum_{i,j,k,l} P(i,j,k,l) \left[\log P(i,j|k,l) - \log P(i|k,l) - \log P(j|k,l)\right] \\
&= \frac{\partial}{\partial P(i|k,l)} - \sum_{j} P(i,j,k,l) \log P(i|k,l) \\
&= \frac{\sum_j P(i,j,k,l)}{P(i|k,l)} \qquad \text{(C.2)}
\end{aligned}$$

The conditional distribution $P(X_t^1|X_{t-1}^1, X_{t-1}^2)$ must fulfill the following constraint:

$$\sum_i P(i|k,l) = 1, \ \forall(k,l) \tag{C.3}$$

which can be forced adding Lagrange multipliers:

$$\lambda\left(\sum_i P(i|k,l) - 1\right) = 0 \tag{C.4}$$

Taking derivatives with respect to $P(i|k,l)$, we obtain:

$$\frac{\partial}{\partial P(i|k,l)}\lambda\left(\sum_i P(i|k,l) - 1\right) = \lambda \tag{C.5}$$

Combining eqs. (C.2) and (C.5), and equaling to 0, we obtain the minimum mutual information transition probabilities for the CMC structure, with respect to the Cartesian product structure:

$$\begin{aligned} \frac{\sum_j P(i,j,k,l)}{P(i|k,l)} + \lambda &= 0 && \text{(C.6)} \\ P(i|k,l) &= \frac{-\sum_j P(i,j,k,l)}{\lambda} && \text{(C.7)} \end{aligned}$$

Since $\lambda$ must normalize the probabilities, the final transition probabilities are given by:

$$\hat{P}(i|k,l) = \frac{\sum_j P(i,j,k,l)}{\sum_i \sum_j P(i,j,k,l)} \tag{C.8}$$

which are the usual transition probabilities computed as frequencies from training data. From the Cartesian product probabilities, these can be obtained by projecting and re-normalizing.

Equivalently, the optimal transition probabilities for $P(X_t^2|X_{t-1}^1, X_{t-1}^2)$ are given by:

$$\hat{P}(j|k,l) = \frac{\sum_i P(i,j,k,l)}{\sum_j \sum_i P(i,j,k,l)} \tag{C.9}$$

# Appendix D

## Expected precision of retrieval with a random selection

This appendix shows how to compute the expected precision of content-based video retrieval when the results are selected randomly. Given the total number of shots in the repository ($N$), the total number of shots retrieved ($M$), and the number of correct shots retrieved ($K$), we ask for the probability of randomly selecting $K$ correct shots within the $M$ shots retrieved. In other words, from all the possible permutations of the total $N$ elements, how many have the correct $K$ elements within a set of $M$.

There are $\binom{M}{K}$ combinations of $K$ elements within a set of $M$. The rest of correct answers $(M-K)$ will be within the remaining $M-N$ elements, with $\binom{M-N}{M-K}$ possible combinations. Furthermore, there are $M!(N-M)!$ different orderings of the correct and incorrect results. Altogether, the probability of randomly selecting $K$ correct answers within the $M$ shots retrieved is:

$$\begin{aligned} P(K) &= \frac{\binom{M}{K}\binom{N-M}{M-K}M!(N-M)!}{N!} && \text{(D.1)} \\ &= \frac{\binom{M}{K}\binom{N-M}{M-K}}{\binom{N}{M}} && \text{(D.2)} \end{aligned}$$

$K$ will usually run from 0 to $M$, as $M$ is the number of true correct results. Therefore, given the usual definition of precision:

$$Pr = \frac{K}{M} \tag{D.3}$$

the expected precision with a random selection of the results would be:

$$Pr_{random} = \sum_{K=0}^{M} \frac{K}{M} P(K) \tag{D.4}$$

However, the case where $M > N - M$ must be taken into account. If there are $M$ true correct results and $M > N - M$, there will always be at least $M - (N - M)$ correct shots retrieved, even with a random selection. Therefore, $K$ will run from $\max(0, 2M - N)$ to $M$, and the final expected precision is:

$$Pr_{random} = \sum_{K=\max(0,2M-N)}^{M} \frac{K}{M} P(K) \tag{D.5}$$

# Appendix E

## Publications

AudiCom is a complete system for monitoring and logging TV commercials. This system is presented in chapter 1 as an example of representing shots by their keyframes, thus turning video segment identification into a still image matching problem. This matching was first addressed using color histograms.

- Juan M. Sánchez and Xavier Binefa. **Automatic digital TV commercial recognition**. In *Proc. VIII National Symposium on Pattern Recognition and Image Analysis*, volume 1, pages 313–320, Bilbao, Spain, May 1999.
- Juan M. Sánchez and Xavier Binefa. **Audicom: a video analysis system for auditing commercial broadcasts**. In *Proc. IEEE Intl. Conf. on Multimedia Computing and Systems*, volume 2, pages 272–276, Firenze, Italy, June 1999.

Then, the problem was addressed using appearance-based matching by PCA dimensionality reduction and Euclidean distance in that low-dimensional space. This a priori simple problem uncovered the main problems found when dealing with digital video. First, a keyframe-based representation of shots relies on the accuracy of the prior shot boundary detection process. Chapter 1 also presents the CECA algorithm, which combines information from color and edges to obtain a robust shot boundary detector.

- Juan M. Sánchez, Xavier Binefa, Jordi Vitrià, and Petia Radeva. **Local color analysis for scene break detection applied to TV commercials recognition**. In *Proc. 3$^{rd}$ Intl. Conf. on Visual Information and Information Systems VISUAL'99*, pages 237–244, Amsterdam, The Netherlands, June 1999. Springer Verlag LNCS 1614.
- Juan M. Sánchez, Xavier Binefa, and Jordi Vitrià. **Shot partitioning based recognition of TV commercials**. *Multimedia Tools and Applications*, 18:233–247, 2002.

The second main problem with digital video combined with color or appearance-based recognition is color appearance variability due to the use of different video acquisition hardware. We developed a GMM based color correction method. It is summarized in chapter 2 and details are given in appendix A. This technique has been applied to appearance-based recognition of keyframes.

- Juan M. Sánchez and Xavier Binefa. **Improving visual recognition using color normalization in digital video applications**. In *Proc. International Conference on Multimedia and Expo*, volume II, pages 1187–1190, New York, NY, July 2000.

These results were compared to other color normalization techniques.

- Juan M. Sánchez and Xavier Binefa. **Color normalization for appearance based recognition of video key-frames**. In *Proc. International Conference on Pattern Recognition*, volume 1, pages 815–818, Barcelona, Spain, September 2000.

And it also has been applied to skin color segmentation for applications like face detection.

- Juan M. Sánchez and Xavier Binefa. **Color normalization for digital video processing**. In *Proc. 4th Intl. Conf. on Visual Information and Information Systems VISUAL'2000*, pages 189–199, Lyon, France, November 2000. Springer Verlag LNCS 1929.

The entire work on GMM color correction has been gathered and extended in

- Juan M. Sánchez and Xavier Binefa. **Video color correction using Gaussian mixture models**. *Submitted to Pattern Recognition*, 2003.

Intermediate-level semantic information must be extracted from low-level features in order to obtain the high-level structure of a video. Our work based on semiotics analyzes the relationships between low-level features and emotional aspects of video production. Mentions to this work are found through the entire dissertation, and particularly in fhapter 2.

- Juan M. Sánchez, Xavier Binefa, Jordi Vitrià, and Petia Radeva. **Linking visual cues and semantic terms under specific digital video domains**. *Journal of Visual Languages and Computing*, 11(3), June 2000.

This work was extended in the Master's thesis.

- Juan M. Sánchez. **Semantic retrieval from digital libraries in the TV commercials domain**. M.Sc. Thesis, CVC Tech. Rep. 29, Centre de Visió per Computador, Universitat Autònoma de Barcelona, September 1999.

Color is known to be an important semantic carrier. In the same way, chapter 2 presents a comprehensive analysis of the semantics conveyed by motion information.

- Juan M. Sánchez and Xavier Binefa. **Semantics from motion in news videos**. In *Proc. IX National Symposium on Pattern Recognition and Image Analysis*, volume 1, pages 79–84, Castellón, Spain, May 2001.

Chapter 3 presents the main contribution of this thesis: the multiple feature temporal modeling of visual contents based on coupled Markov chains. This chapter is mainly focused on the definition of the model and the development of a method for automatic structure learning in terms of minimum cost and maximum representation accuracy.

- Juan M. Sánchez, Xavier Binefa, and John R. Kender. **Combining the representation of multiple features in temporal models for the characterization of visual contents in video**. In *Proc. International Conference on Image and Video Retrieval (CIVR'2003)*, pages 216–226, Urbana, IL, July 2003. Springer Verlag LNCS 2728.

Different applications of the CMC modeling to the semantic analysis of videos were presented in chapter 4. The CMC representation provides semantically meaningful clusterings of video shots, which can be combined with simple domain knowledge to obtain the high-level structure of News videos.

- Juan M. Sánchez, Xavier Binefa, and John R. Kender. **Coupled Markov chains for video contents characterization**. In *Proc. International Conference on Pattern Recognition*, Quebec, Canada, August 2002.

Object detection and localization in video sequences is another application of the CMC representation.

- Juan M. Sánchez, Xavier Binefa, and John R. Kender. **Multiple feature temporal models for object detection in video**. In *Proc. IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, August 2002.

During the process of obtaining the CMC representations of the shots of a video, shot boundaries are detected using the same model and similarity measures. This process can also be used as a stand-alone shot boundary detector.

- Juan M. Sánchez and Xavier Binefa. **Shot segmentation using a coupled Markov chains representation of video contents**. In *Proc. 1st. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'03)*, pages 902–909, Mallorca, Spain, June 2003. Springer Verlag LNCS 2652.

# Bibliography

[1] D.A. Adjeroh, M.C. Lee, and I. King. A distance measure for video sequence similarity matching. In *Proc. Int. Workshop on Multimedia Database Management Systems*, pages 72–79, Dayton, OH, August 1998.

[2] Philippe Aigrain, Philippe Joly, and Véronique Longueville. Medium knowledge-based macro-segmentation of video into sequences. In M.T. Maybury, editor, *Intelligent Multimedia Information Retrieval*, pages 159–173. AAAI/MIT Press, 1997.

[3] Aya Aner. *Video summaries and cross-referencing.* PhD thesis, Columbia University, July 2002.

[4] Lluis Barceló and Xavier Binefa. Bayesian video mosaicing with moving objects. In *Proc. IX National Symposium on Pattern Recognition and Image Analysis*, volume 1, pages 91–96, Castellón, Spain, May 2001.

[5] Ana Benitez and Shih-Fu Chang. Multimedia knowledge integration, summarization and evaluation. In *Proc. Intl. Workshop on Multimedia Data Mining*, Edmonton, Canada, July 2002.

[6] Michael J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.

[7] John S. Boreczky and Lawrence A. Rowe. Comparison of video shot boundary detection techniques. In I.K. Sethi and R.C. Jain, editors, *Proc. Storage and Retrieval for Image and Video Databases IV, SPIE*, volume 2670, pages 170–179, 1996.

[8] Patrick Bouthemy and Ronan Fablet. Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *Proc. International Conference on Pattern Recognition*, volume 1, pages 905–908, Brisbane, Australia, August 1998.

[9] David H. Brainard and William T. Freeman. Bayesian color constancy. *Journal of the Optical Society of America A*, 14(7):1393–1411, July 1997.

[10] Marco Bressan, David Guillamet, and Jordi Vitrià. Using an ICA representation of local color histograms for object recognition. *Pattern Recognition*, 36(3):691–701, 2003.

[11] M. Caliani, C. Colombo, A. Del Bimbo, and P. Pala. Commercials video retrieval by induced semantics. In *Proc. IEEE Workshop on Content-based Access of Image and Video Databases*, pages 72–80, Bombay, India, January 1998.

[12] Shih-Fu Chang and Hari Sundaram. Structural and semantic analysis of video. In *Proc. IEEE Int. Conf. on Multimedia and Expo*, New York, NY, July 2000.

[13] Carlo Colombo, Alberto Del Bimbo, and Pietro Pala. Retrieval of commercials by semantic content: the semiotic perspective. *Multimedia Tool and Applications*, 13:93–118, 2001.

[14] J.M. Corridoni, A. Del Bimbo, and P. Pala. Sensations and psychological effects in color image databases. *ACM Multimedia Systems Journal*, 7(3):175–183, 1999.

[15] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1977.

[16] Stefan Eickeler and Stefan Müller. Content-based video indexing of TV broadcast news using hidden Markov models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2997–3000, Phoenix, AZ, March 1999.

[17] Ronan Fablet and Patrick Bouthemy. Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval. In *Proc. 3rd Int. Conf. on Visual Information Systems VISUAL'99*, pages 221–228, Amsterdam, The Netherlands, June 1999.

[18] Ronan Fablet, Patrick Bouthemy, and Patrick Pérez. Non-parametric statistical analysis of scene activity for motion-based video indexing and retrieval. *IRISA Technical Report 1351*, September 2000.

[19] Mark D. Fairchild. *Color Appearance Models.* Addison-Wesley, Reading, MA, 1998.

[20] G. Finlayson, M. Drew, and B. Funt. Colour constancy: Generalized diagonal transforms suffice. *Journal of the Optical Society of America A*, 11(11):3011–3020, 1994.

[21] G. Finlayson, B. Funt, and J. Barnard. Colour constancy under a varying illumination. In *Proc. Fifth Intl. Conf. on Computer Vision*, June 1995.

[22] G.D. Finlayson, B. Schiele, and J.L. Crowley. Comprehensive colour image normalization. In Hans Burkhardt and Bernd Neumann, editors, *Proc. of 5th European Conference on Computer Vision*, volume I, pages 475–490, Freiburg, Germany, 1998.

[23] B.V. Funt and G.D. Finlayson. Color constant color indexing. *IEEE Trans. Patt. Anal. and Mach. Intell*, 17(5), May 1995.

[24] Ullas Gargi, Rangachar Kasturi, and Susan H. Strayer. Performance characterization of video-shot-change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1), February 2000.

[25] Daniel Gatica-Perez, Alexander Loui, and Ming-Ting Sun. Finding structure in consumer videos by probabilistic hierarchical clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2003.

[26] Theo Gevers and A.W.M. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.

[27] Theo Gevers and H.M.G. Stokman. Robust histogram construction from color invariants for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10), 2003.

[28] Louis Giannetti. *Understanding Movies (8th edition).* Prentice Hall, 1999.

[29] Andreas Girgensohn and John Boreczky. Time-constrained keyframe selection technique. *Multimedia Tools and Applications*, 11:347–358, 2000.

[30] Yihong Gong and Xin Liu. Generating optimal video summaries. In *IEEE International Conference on Multimedia and Expo*, pages 1559–1562, New York, NY, July 2000.

[31] Monika M. Gorkani and Rosalind W. Picard. Texture orientation for sorting photos “at a glance”. In *Proc. International Conference on Pattern Recognition*, volume A, pages 459–464, Jerusalem, Israel, October 1994.

[32] P.O. Gresle and T.S. Huang. Gisting of video documents: A key frames selection algorithm using relative activity measure. In *The 2nd International Conference on Visual Information Systems*, 1997.

[33] Arun Hampapur. Semantic video indexing: Approach and issues. *SIGMOD Record*, 28(1):32–39, 1999.

[34] Arun Hampapur and Ruud Bolle. Comparison of distance measures for video copy detection. In *Proc. IEEE Int. Conf. on Multimedia and Expo*, Tokyo, Japan, August 2001.

[35] R. Haralick. Statistical and structural approaches to texture. *Proc. IEEE*, 67:786–804, May 1979.

[36] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Proc. IEEE Computer Vision and Pattern Recognition Conference*, CVPR’97, San Juan, Puerto Rico, June 1997.

[37] International Color Consortium (ICC). File format for color profiles. Specification ICC.1:2001-12. http://www.color.org, 2002.

[38] Internet Movie Database. http://www.imdb.com.

[39] Giridharan Iyengar and Andrew B. Lippman. Content-based browsing and editing of unstructured video. In *Proc. IEEE Int. Conf. on Multimedia and Expo*, volume 1, pages 159–162, New York, NY, 2000.

[40] Finn V. Jensen. *An introduction to Bayesian Networks*. UCL Press, 1996.

[41] John R. Kender and Boon-Lock Yeo. Video scene segmentation via continuous video coherence. In *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*, pages 367–373, June 1998.

[42] Irena Koprinska and Sergio Carrato. Temporal video segmentation: A survey. *Signal Processing: Image Communication*, 16:477–500, 2001.

[43] Riccardo Leonardi and Pierangelo Migliorati. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9(2):44–51, 2002.

[44] R. Lienhart, C. Kuhmünch, and W. Effelsberg. On the detection and recognition of television commercials. In *Proc. IEEE Conf. on Multimedia Computing and Systems*, pages 509–516, Ottawa, Canada, June 1997.

[45] Tiecheng Liu and John R. Kender. A hidden Markov model approach to the structure of documentaries. In *Proc. Workshop on Content Based Access to Image and Video Libraries*, pages 111–115, Hilton Head, SC, June 2000.

[46] Tiecheng Liu and John R. Kender. On the structure and analysis of home videos. In *Proc. Asian Conference on Computer Vision*, 2000.

[47] Joan Llach and Philippe Salembier. Analysis of video sequences: table of contents and index creation. In *Proc. Intl. Workshop on Very Low Bitrate Video VLBV'99*, pages 111–115, Kyoto, Japan, October 1999.

[48] C.S. McCamy, H. Marcus, and J.G. Davidson. A color-rendition chart. *Journal of Applied Photographic Engineering*, 2(3):95–99, Summer 1976.

[49] Bartlett W. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

[50] Movies.com. http://movies.go.com.

[51] H. Murase and K. Nayar. Visual learning and recognition of 3-D objects from appearance. *Int. Journal of Computer Vision*, 14(1):5–24, 1995.

[52] Milind R. Naphade, Roy Wang, and Thomas S. Huang. Multimodal pattern matching for audio-visual query and retrieval. In *Proc. SPIE Storage and Retrieval for Media Databases*, volume 4315, San Jose, CA, January 2001.

[53] Milind R. Naphade, Roy Wang, and Thomas S. Huang. Supporting audiovisual query using dynamic programming. In *Proc. ACM Multimedia Conference*, pages 411–420, Ottawa, Canada, October 2001.

[54] Randal C. Nelson and Ramprasad Polana. Qualitative recognition of motion using temporal texture. *CVGIP: Image Understanding*, 56(1):78–79, July 1992.

[55] Chong-Wah Ngo, Ting-Chuen Ponh, and Hong-Jiang Zhang. On clustering and retrieval of video shots. In *Proc. ACM Multimedia Conference*, Ottawa, Canada, October 2001.

[56] Xavier Orriols and Xavier Binefa. An EM algorithm for video summarization, generative model approach. In *IEEE International Conference on Computer Vision*, volume 2, pages 335–342, Vancouver, Canada, July 2001.

[57] Colin O'Toole, Alan Smeaton, Noel Murphy, and Sean Marlow. Evaluation of automatic shot boundary detection on a large video test suite. In *Proc. Challenge of Image Retrieval, 2nd UK Conf. on Image Retrieval (CIR'99)*, Newcastle, UK, February 1999.

[58] Seungyup Paek and Shih-Fu Chang. A knowledge engineering approach for image classification based on probabilistic reasoning systems. In *Proc. IEEE Int. Conf. on Multimedia and Expo*, volume 2, pages 1133–1136, New York, NY, July 2000.

[59] Constantine P. Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Proc. International Conference on Computer Vision*, pages 555–562, Bombay, India, January 1998.

[60] Constantine P. Papageorgiou and Tomaso Poggio. A pattern classification approach to dynamical object detection. In *Proc. International Conference on Computer Vision*, pages 1223–1228, Corfu, Greece, September 1999.

[61] Greg Pass and Ramin Zabih. Comparing images using joint histograms. *Multimedia Systems*, 7(3):234–240, 1999.

[62] Greg Pass, Ramin Zabih, and Justin Miller. Comparing images using color coherence vectors. In *ACM Multimedia*, pages 65–73, 1996.

[63] Kadir A. Peker, A. Aydin Alatan, and Ali N. Akansu. Low-level motion activity features for semantic characterization of video. In *Proc. IEEE International Conference on Multimedia and Expo*, volume II, pages 801–804, New York, NY, July 2000.

[64] R.W. Picard and T.P. Minka. Vision texture for annotation. *Multimedia Systems*, 3(3):3–14, February 1995.

[65] Richard Qian, Niels Haering, and Ibrahim Sezan. A computational approach to semantic event detection. In *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 1200–1206, Fort Collins, CO, June 1999.

[66] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Constructing Table-of-Content for videos. *ACM Multimedia Systems Journal*, 7(5):359–368, September 1999.

[67] Erez Sali and Shimon Ullman. Combining class-specific fragments for object classification. In *Proc. British Machine Vision Conference*, pages 203–213, Nottingham, UK, September 1999.

[68] Andrea Selinger and Randal C. Nelson. Improving appearance-based object recognition in cluttered backgrounds. In *Proc. International Conference on Pattern Recognition*, volume 1, pages 46–50, Barcelona, Spain, September 2000.

[69] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[70] Martin Szummer and Rosalind W. Picard. Temporal texture modeling. In *Proc. IEEE International Conference on Image Processing*, volume 3, pages 823–826, Lausanne, Switzerland, September 1996.

[71] Martin Szummer and Rosalind W. Picard. Indoor-outdoor image classification. In *Proc. Workshop on Content-Based Access of Image and Video Databases*, pages 42–51, Bombay, India, January 1998.

[72] Xavier Ubiergo Cabedo and Sushil K. Bhattacharjee. Shot detection tools in digital video. In *Proc. Noblesse Workshop on Non-linear Model Based Image Analysis (NMBIA'98)*, pages 231–236, Glasgow, Scotland, July 1998.

[73] A. Vailaya, M. Figueiredo, A. Jain, and Hong Jiang Zhang. Content-based hierarchical classification of vacation images. In *Proc. International Conference on Multimedia Computing and Systems*, volume 1, pages 518–523, Florence, Italy, June 1999.

[74] Aditya Vailaya, Anil Jain, and HongJiang Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935, 1998.

[75] Aditya Vailaya, Anil K. Jain, and HongJiang Zhang. Video clustering. Technical Report MSU-CPS-96-64, Department of Computer Science, Michigan State University, East Lansing, Michigan, November 1996.

[76] Nuno Vasconcelos and Andrew Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Transactions on Image Processing*, 9(1):3–19, January 2000.

[77] Emmanuel Veneau, Rémi Ronfard, and Patrick Bouthemy. From video shot clustering to sequence segmentation. In *Proc. International Conference on Pattern Recognition*, Barcelona, Spain, September 2000.

[78] Paul Viola and Michael Jones. Robust real-time object detection. In *2nd Intl. Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, July 2001.

[79] W. Wolf. Key frame selection by motion analysis. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, 1996.

[80] Minerva M. Yeung and Bede Liu. Efficient matching and clustering of video shots. In *Proc. Intl. Conf. on Image Processing*, volume 1, pages 338–341, 1995.

[81] Minerva M. Yeung and Boon-Lock Yeo. Time-constrained clustering for segmentation of video into story units. In *Proc. Intl. Conf. on Pattern Recognition*, volume C, pages 375–380, August 1996.

[82] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *ACM Conference on Multimedia*, San Francisco, California, November 1995.

[83] H. Zhang, J. Wu, D. Zhong, and S.W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.

[84] H. J. Zhang, A. Kankanhalli, and S. Smoliar. Automatic partitioning of video. *Multimedia Systems*, 1(1):10–28, 1993.

[85] Di Zhong and Shih-Fu Chang. Structure analysis of sports videos using domain models. In *Proc. IEEE Int. Conf. on Multimedia and Expo*, Tokyo, Japan, August 2001.

[86] Di Zhong, HongJiang Zhang, and Shih-Fu Chang. Clustering methods for video browsing and annotation. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pages 239–246, 1996.

[87] Yueting Zhuang, Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *IEEE International Conference on Image Processing*, pages 866–870, Chicago, IL, October 1998.