Chapter 2

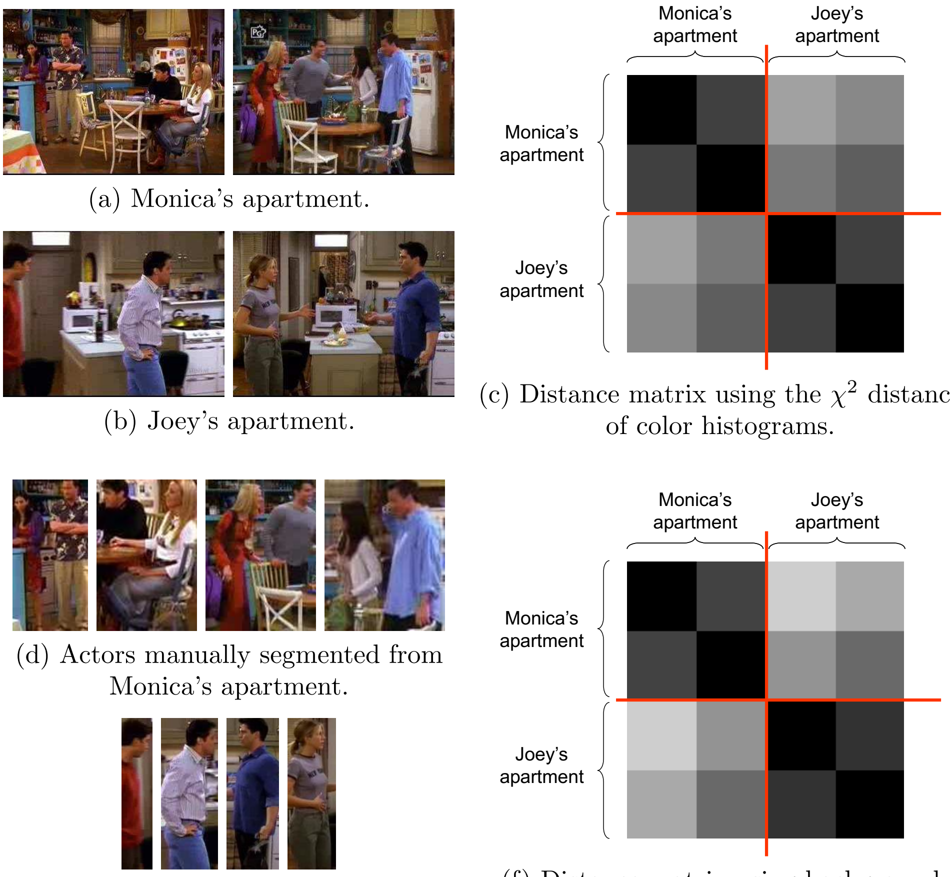
Semantics from low-level features

Objects, background and other elements that form a scene have their own characteristic low-level features. A person standing in front of a camera will be the cause of certain colors and textures that appear in the image sequence that is being created. If that person walks, a certain motion pattern will be observed as well. Image features carry information that characterizes the semantic concepts that caused them. This chapter explores the semantic information that is implicitly contained in low-level features. This information can be used for video indexing and annotation, and also to obtain intermediate-level semantic descriptions of contents to be used for higher level video structure analysis. Color and motion are two main low-level features that can be used for these purposes. First, the use of color as a semantic carrier is reviewed. Then, the semantics that can be inferred from motion information is analyzed in the domain of news videos. Finally, different ways of combining multiple features for the characterization of semantic concepts are reviewed as well.

2.1 Semantics from color

Color has been used as the simplest way to do object recognition and retrieval since the introduction of color histogramming techniques for video indexing by Swain and Ballard in [69]. A color histogram summarizes the colors of an object. Therefore, certain information about the appearance of an object and its identity is contained in it. A color histogram can thus be seen as a kind of intermediate-level semantic representation of an object.

When the object that is characterized using color turns out to be the background of the scene, a characterization of the location is obtained. For example, the images in fig. 2.1(a,b) were shot at two different locations from the sitcom Friends: Monica's and Joey's apartments. Color histograms and the χ^2 distance were used to compute the distance matrix shown in fig. 2.1(c), where the two clusters can be clearly noticed.



(a) Monica's apartment.

(b) Joey's apartment.

(c) Distance matrix using the χ^2 distance of color histograms.

(d) Actors manually segmented from Monica's apartment.

(e) Actors manually segmented from Joey's apartment.

(f) Distance matrix using background segmentation.

Figure 2.1: Clustering shots by location using the color of the background.

Note that the color histograms were computed for the full images, without prior foreground/background segmentation. This segmentation could yield much better results, as the color histograms would only consider the appearance of the background. For instance, figs. 2.1(d,e) show the actors that appeared in each scene, which were manually segmented in this case. Figure 2.1(f) shows the new distance matrix after the actors had been removed from the images. Aner developed in [3] a method for indexing videos by location based on this idea. She used mosaics instead of simple keyframes, so that the whole portion of background captured during a shot was considered. At the same time, moving objects were removed from the scenes. She then performed a color-based comparison of the mosaics.

This idea has also been used for clustering video shots in terms of their location in order to obtain a segmentation of the story units that compose a video. This approach is followed by Yeung and Yeo in [81]. They first cluster video shots using the method

Color	Feelings evoked in the viewer
Red	Happiness, dynamism, aggressiveness, violence, power
Orange	Glory, solemnity, vanity, progress
Golden yellow	Richness, prosperity, happiness
Dark yellow	Deception, caution
Green	Calm, relax, hope
Blue	Gentleness, fairness, faithfulness, virtue
Purple	Melancholy, fear
Brown	Relax (mostly used as background color)

Table 2.1: Feelings related to colors, as considered by semiotics.

from [80], which is based on the RGB color histogram intersection distance. Higher-level knowledge about the production process in the domain of sitcoms is then used to generate a Scene Transition Graph (STG). The main assumption is that repeated shots of the same persons or same settings, alternating or interleaving with other shots, are often deployed in many programs to convey parallel events in a scene, such as conversations and reactions. Temporal constraints are also applied, so that if two visually similar shots occur far apart in the time line, they may potentially belong to different scenes. A similar approach is followed in [41] by Kender and Yeo. They measure probable scene boundaries by calculating a short term memory-based model of shot-to-shot coherence. The assumption that visually similar shots repeat in a scene and the temporal constraint are also considered.

A different point of view about the semantics conveyed by color is given from a semiotics perspective. Semiotics is concerned with unspoken messages that are communicated to the observer by the use of visual features. For instance, warm colors grab the attention of the viewer and convey dynamism. On the other hand, cold colors suggest gentleness, calm, relax and faithfulness. Each color can be associated to a set of feelings, which are summarized in table 2.1. The use of saturated colors is considered a sign of unrealistic situations, giving a sense of fancy and joyful worlds, and thus communicates happiness. The presence of light colors also induces the viewer to feel calm and relax. All these observations can be used to organize a video archive in terms of complex semantic concepts.

Two main ways of improving the performance of color descriptions for object recognition and retrieval have been proposed and are reviewed next.

2.1.1 Invariant color representations

These works are concerned on obtaining color features that do not depend on some specific image formation factors like lightning conditions. The choice of color space may result in some kind of invariance. For example, some authors prefer to use the HSV color space and drop the V component, which is directly related to luminance, instead of the RGB space, where the three components are affected by luminance. The HSV color space is depicted in fig. 2.2. The transformation from RGB values to

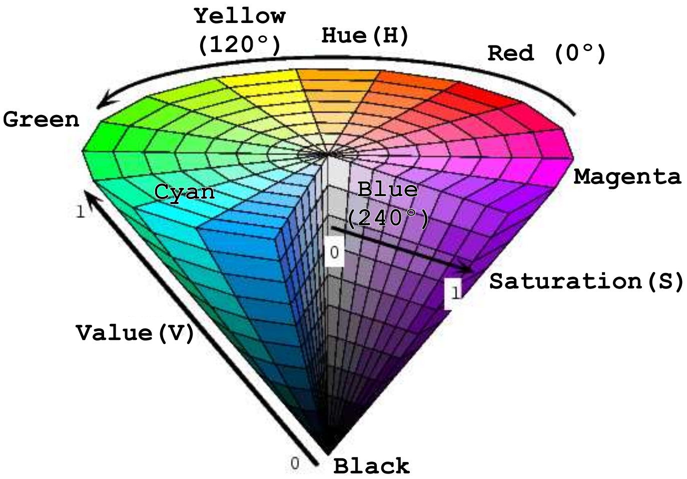


Figure 2.2: The HSV color space.

HSV is given by the following equations:

$$V = max\{R, G, B\} \tag{2.1}$$

$$\Delta = V - min\{R, G, B\}$$

$$S = \frac{\Delta}{V} \tag{2.2}$$

$$H = \begin{cases} \frac{\pi}{3}\left(\frac{G-B}{\Delta}\right) & if \quad R = \Delta \\ \frac{\pi}{3}\left(2 + \frac{B-R}{\Delta}\right) & if \quad G = \Delta \\ \frac{\pi}{3}\left(4 + \frac{R-G}{\Delta}\right) & if \quad B = \Delta \end{cases} \tag{2.3}$$

Gevers and Smeulders proposed in [26] color models that are invariant to changes in viewing direction, object geometry, illumination and highlights. They report object recognition results on a database of 70 objects and 500 images, and they analyze the performance of different color models under variations of these image formation parameters. One significant conclusion of their work is that the RGB color model has the worst performance due to its sensitivity to varying image conditions.

A different approach is computational color constancy, which aims to obtain color representations that depend only on spectral surface reflectances and not on the illumination. Different techniques have been proposed by Brainard and Freeman [9], Finlayson et al. [22, 21] and Funt and Finlayson [23], amongst others. Although it is known that color constancy does not exist in the human visual system, the computational approach produces models that are closely related to the chromatic adaptation process [19].

In the context of digital video libraries, color appearance variations can also be caused by factors that cannot be strictly considered as image formation parameters.

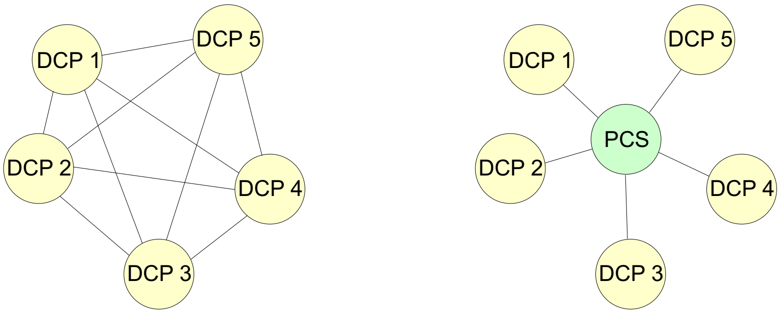


Figure 2.3: In general, for N device-dependent color profiles (DCP), we would need $N(N-1)$ color space transformations (left). The ICC defines a device-independent color space, called Profile Connection Space (PCS), so that only one transform per device is needed (right).

The most typical case is color appearance variations due to the use of different acquisition hardware, where the sensitivity of their sensors may vary. In this case, color invariants and color constancy algorithms may not be appropriate. A more suitable approach for this kind of variations is finding mappings between device-dependent color spaces. The International Color Consortium (ICC) has made an effort to standardize device-dependent color profiles [37]. This standard defines a Profile Connection Space (PCS), which is a device-independent color space, so that only one transformation must be defined per input or output device. This scheme is depicted in fig. 2.3. The ICC work has been specifically oriented to desktop publishing applications. Its application is arguable in the domain of digital video libraries, where the source of a video may be unknown, and the footage goes through several color transformations caused by the VCR, the frame grabber, and other hardware that may be involved in the digitization process.

Device-dependent color appearances can be characterized by the parameters of a set of Gaussian distributions. In this way, the intrinsic appearance of a color is determined by the contribution of each Gaussian distribution to it. Mappings are then defined between different device-dependent color spaces in order to keep the intrinsic appearance of colors, that is, their identities. A device-independent color appearance space is also defined as a normalized representation of color identities. In this way, the ICC approach can be implemented, and only one mapping per device must be defined. They report experimental results on applications like skin color segmentation and image retrieval, and a comparison with color constancy approaches. Their experiments show that the grayworld color constancy algorithm provides as good results as their method, but with a lower computational cost. However, color appearances are not preserved using the grayworld approach, so that it cannot be used for the skin color segmentation application. Therefore, the conclusion is inverted: the Gaussian mixture approach to color correction is appropriate for applications where color appearance must be conserved, and its performance is high for image retrieval purposes as well. The main disadvantage of this method is that a calibration pattern, like the ones in fig. A.5, has to be acquired using the hardware that is characterized, and this may not always be possible. More details about the characterization of device color spaces, and mappings between them are given in appendix A.

2.1.2 Enhanced color representations

Different extensions to color histograms have been developed. They are mainly concerned on including spatial information within the summarized color representation in order to capture the shape and distribution of the colors of the object or image. Color histograms lack spatial information, and this can cause images with very different appearances to have very similar histograms. Color coherence vectors (CCV) were defined by Pass et al. in [62]. The CCV measures the spatial coherence of the pixels of a given color. If regions of a certain color in the image are large, then that color has high coherence, and has low coherence otherwise. Huang et al. introduced the color correlogram in [36]. They define the color correlogram as a table indexed by color pairs, where the k-th entry for $< i, j >$ specifies the probability of finding a pixel of color j at a distance k from a pixel of color i in the image. They conclude that such an image feature is robust in tolerating large changes in appearance of the same scene caused by changes in viewpoint positions, changes in the background scene, partial occlusions and camera zoom that causes radical changes in shape. The definition is basically the same as that of a cooccurrence matrix for the representation of graylevel spatial textures. Therefore, the color correlogram is a way of characterizing a statistical color texture.

2.2 Semantics from motion

Motion information analysis usually requires the segmentation of different moving objects and background entities. This task is particularly challenging on its own and is matter of deep research [4]. On the other hand, motion patterns can be represented using temporal motion textures. Temporal motion textures extend classical grayscale texture analysis techniques. The idea is to characterize patterns of motion along time. These patterns, like in spatial textures, can be either statistical (windblown trees) or structural (a person walking). In contrast, we find motion events, which are single events that do not repeat in space or time (opening a door).

Nelson and Polana showed in [54] that certain statistical spatial and temporal features that can be derived from approximations to the motion field have invariant properties, and can be used to classify regional activities such as windblown trees or chaotic fluid flow, that are characterized by complex, non-rigid motion. They used a set of statistical features computed on the normal flow field for each texture in order to characterize and classify the following textures: fluttering crepe paper bands, cloth waving in the wind, motion of tree in the wind, flow of water in a river, turbulent motion of water, uniformly expanding image produced by forward observer motion, and uniformly rotating image produced by observer roll. Szummer and Picard modeled temporal textures in [70] using the spatio-temporal autoregressive model (STAR), which expresses each pixel as a linear combination of surrounding pixels lagged both in space and in time. This model not only provides a base for recognition, but also for synthesis of temporal textures.

A statistical characterization of a temporal motion texture will globally consider motion information from the whole scene, thus including global and object motions in it. In this way, object/background segmentation is not required to represent the motion patterns in a video shot. A representation of temporal motion textures based on their temporal cooccurrence matrices is presented next. This method captures the underlying motions in the scene, as well as their temporal variations.

2.2.1 Temporal motion texture modeling

In the same way a spatial texture is regarded as a particular spatial distribution of gray level values, a temporal motion texture can be seen as a distribution of spatio-temporal motion measures. Bouthemy and Fablet's approach [8] is based on extending the well-known characterization of textures using cooccurrence matrices developed by Haralick in [35]. For a spatial texture, each value $P_d(i,j)$ in the cooccurrence matrix P_d contains the probability of finding the values i and j with a spatial distance d in the texture. The extension to spatio-temporal motion distributions is straightforward, being $P_d(i,j)$ the probability of finding motion observations i and j at a temporal distance d, and in the same spatial position. We will consider the norm of the velocity vectors as our observations. These observations must be considered along a significant set of frames in order to correctly capture the temporal behavior of the texture. Thus, the temporal cooccurrence for the pair of motion observations (i,j) at the temporal distance d in image sequence $I(x,y,t)$, $t \in [t_1,t_2]$ is defined as:

$$P_d(i,j) = \frac{\#\{(x,y,t) \mid v_{obs}(x,y,t) = i,\ v_{obs}(x,y,t+d) = j,\ t,(t+d) \in [t_1,t_2]\}}{\#\{(x,y,t) \mid t,(t+d) \in [t_1,t_2]\}} \tag{2.4}$$

where $v_{obs}(x,y,t)$ is the motion observation in position (x,y) and time t.

A reduced set of statistical descriptors can then be obtained from a cooccurrence matrix in order to obtain a reduced and meaningful characterization of the underlying temporal texture. Bouthemy and Fablet report two sets of these descriptors:

1. Entropy, inverse difference moment, acceleration, kurtosis and difference kurtosis.

2. Average, variance, Dirac, angular second moment (ASM) and contrast.

These descriptors are respectively defined in [17] and [8]. The second set has the advantage that each feature has an interpretation in terms of motion perception. The average is directly related to the amount of motion, whereas variance and Dirac show the degree of spreading of the motion distribution. The ASM measures the temporal coherence of motion and contrast is related to the average acceleration. These descriptors are mathematically defined as:

- Average: $A = \sum_{(i,j)} iP_d(i,j)$
- Variance: $\sigma^2 = \sum_{(i,j)} (i-A)^2 P_d(i,j)$
- Dirac: $\delta = A^2/\sigma^2$
- Angular Second Moment: $ASM = \sum_{(i,j)} P_d(i,j)^2$
- Contrast: $Cont = \sum_{(i,j)} (i-j)^2 P_d(i,j)$

In order to compute the cooccurrence matrix, motion observations must be quantized. The norm of the velocity vectors at each spatial location is considered. Note that the orientation component of velocity is dropped. Quantization is a delicate step, as the dynamic range of motion observations has to be taken into account. This range is domain-dependant, and in the case of news videos the maximum motion found is commonly small. In this case, 16 quantization levels within the range [0,3] is a suitable value, so that cooccurrence matrices P_d will be sized 16×16.

It is important to note that the accuracy needed when computing velocity vector fields is not necessarily high. Noise and computation errors at this level will not have a significant effect on the final descriptors that will characterize a temporal texture. We used an accurate algorithm for optical flow computation by Black and Anandan [6], which embeds previous common approaches within a robust estimation framework. This approach takes into account possible violations of the data conservation and spatial coherence constraints on image motion. These constraints are necessary to make optical flow computation a well-posed problem, but can lead to estimation errors when they are not completely fulfilled. However, tests performed using a simple correlation method for optical flow estimation show that the final descriptors obtained are practically the same.

2.2.2 Semantic classification based on temporal motion texture

We have seen that the automatic detection of anchors in news videos plays a key role in order to obtain their high-level semantic structure. Besides, special correspondents and people relevant to the piece of news, i.e. politicians or other celebrities, are significant in the indexing and annotation senses. This section is focused on finding shots of individuals using motion information, considering that these shots appear as close-ups and medium shots as they are defined by film-making terminology [28]. The difference between a close-up and a medium shot is basically defined by the distance from the camera to the subject matter. Considering a person shot, a close-up will show mainly his/her face or head, while a medium shot would include head, chest and arms. Examples are shown in fig. 2.4.

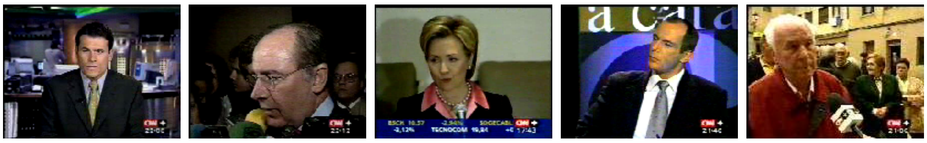


Figure 2.4: Close-ups and medium shots containing individuals.

Peker et al. observed in [63] two main facts that can be used as heuristics for detecting close-ups: low coherence of motion along time and relatively large motions. Both facts are due to the short distance between the camera and the object. The following observations on the motion-related descriptors obtained from our data set are common for close-up shots with significant motion:

- relatively high average measure, expressing large motions,
- high variance and Dirac, expressing sparsity of motion cooccurrences,
- low ASM, showing low temporal coherence of motion,
- and high contrast, which is related to a high average acceleration given by sudden motions.

These observations are fully consistent with the previously discussed heuristics, as they are expressing the presence of non-coherent significant motions. However, medium shots do not fulfill these requirements, as they are basically shots with very little motion due to the bigger distance between camera and object. Both kinds of shots should be included in a class of “1-person shots”.

Feature descriptors were computed on 342 shots from a set of news videos, thus obtaining a representation of the original data in a 5-dimensional feature space. 152 of them were labeled as “1-person shots” and 190 were labeled as “other”. Principal Component Analysis on this data showed up high correlations, as over 99% of the total variance of the original data was kept in a 2-dimensional subspace spanned by their principal axis. The highest coefficients in the linear combination correspond to variance and contrast features. Coefficients for average are lower, and those corresponding to Dirac and ASM are practically 0. Besides, this dimensionality reduction will allow us to spatially observe and analyze the data distributions.

Given the different classes C_n defined by our classification problem (“1-person shot” vs. “other”), their distributions in feature space can be modeled as likelihood functions $P(x|C_n)$ in order to use a Bayesian classifier in the experiments. In this framework, a shot is assigned to the class C_i that satisfies:

$$P(C_i|x) > P(C_j|x), \ \ \forall C_j \neq C_i \tag{2.5}$$

where x is the vector of feature descriptors of the shot and $P(C_n|x)$ is defined by Bayes’ rule as:

$$P(C_n|x) = \frac{P(x|C_n)P(C_n)}{P(x)} \tag{2.6}$$

For this particular classification problem, the two classes defined stand for whether a shot contains a person speaking to the camera or not ("1-person shots" and "others"). Figure 2.4 shows the wide variety of camera shot distances and orientations that were considered as "1-person shots" in the experiments. To cope with complex distributions, the probability density of each class $P(x|C_n)$ is assumed to follow a Gaussian mixture model. A key parameter of this kind of distributions is the number of Gaussian components in it. The number of components is automatically selected using a Minimum Description Length (MDL) criterion. The other parameters of each distribution are estimated from data using the EM algorithm. The prior probabilities for each class $P(C_n)$ can either be assigned all the same value, assuming no prior knowledge, or be computed from the relative frequency of each class, so that the most observed class is the most probable. Finally, the unconditional probability of observation x is given by:

$$P(x) = \sum_{\forall n} P(x|C_n)P(C_n) \tag{2.7}$$

The classifier can be directly applied to the samples in the original 5-dimensional feature space. However, PCA suggested a high correlation between features in the original space, so that either the original feature space or the one spanned by their principal components can be used. The second one is preferred in order to reduce the computational cost and to obtain better estimates of the Gaussian mixture distribution parameters. Figure 2.5 shows the distribution of the samples of the two classes in this subspace. The contour plots correspond to the Gaussian mixture estimates for each data set. We can see that the class of "1-person shots" is mainly concentrated in a very definite region of space, but two Gaussian components where still required in order to properly characterize this class. This is in keeping with the fact that close-ups and medium shots, which have different motion characteristics, have been considered in the same class. Note that most of the shots were medium shots, which means that the Gaussian component of the mixture that corresponds to medium shots has higher density. On the other hand, the elements of the "others" class are much more sparse.

The common strategy for evaluating the performance of a classifier is based on defining a training and a test data sets, which are respectively used to estimate the parameters of each class distributions and to evaluate them. However, in some cases, there are not enough data samples available to divide them into populated data sets that will allow us to obtain good parameter estimates and significant evaluation measures. In these cases, the leave-one-out strategy is known to provide results as significant as those obtained using dense training and test data sets, at the cost of a very computationally expensive process. Classification results obtained using this strategy are shown as a confusion matrix in table 2.2. The total correct classification rate obtained was 77.63%.

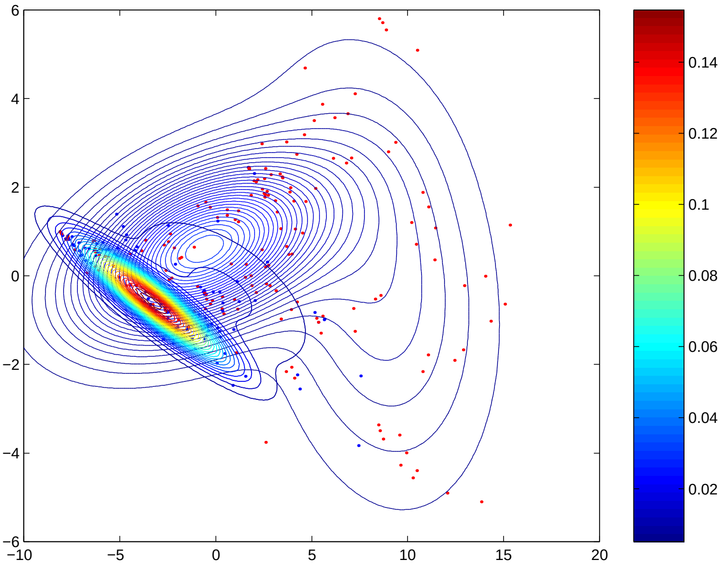


Figure 2.5: Gaussian mixture distributions for the classes “1-person shots” and “other shots”.

	Classified as		
	1-person shot	Other	Total
1-person shots	128	24	152
Others	55	135	190

Table 2.2: Confusion matrix of the “1-person shots” classifier.

Most misclassifications are due to wrongly assigning shots to the “1-person shots” class. Some of them are shown in fig. 2.6. However, it is interesting to note that many of them are medium shots of two or three people, known in film-making as two-shots and three-shots. Their motion activity pattern is basically the same as in a single-person medium shot. This was the case in 21 out of 55 misclassifications. We can also observe close-up shots where the subject matter is not a person, like a hand writing on a paper or a waving flag. Figure 2.6 also shows a close-up into a crowd. The motion texture patterns found in these shots are certainly very similar to the ones in the class of “1-person shots”, as they basically depend on the distance from the camera to the object, and not on the type of object itself. The rest of misclassified elements of the class “others” show a general low motion activity pattern, so that they can be mistaken as medium shots when only motion-based features are considered. After these observations, the correct classification rate obtained can be considered successful.

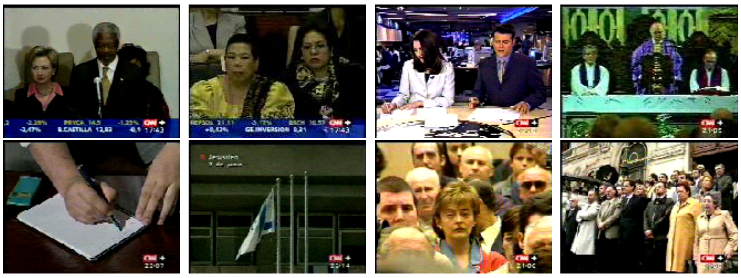


Figure 2.6: "Other shots" wrongly classified as "1-person shots".

2.2.3 Discussion

Temporal motion textures provide semantically meaningful information about the visual contents in the shot. The initial approach in this section aimed to classify one-person shots using a representation of their temporal motion texture based on temporal motion cooccurrences. However, experimental results show that the set of features selected for classification are mainly related to the type of shot, basically close-up, medium or long shot, and not to the specific subject matter that is filmed. The high correlation found between the features used suggests that some of them could be superfluous. Other descriptors computed on cooccurrence matrix values could also work as hidden variables, providing meaningful information related to non-obvious characteristics of data. This work suggests that other semantically meaningful interpretations of motion-related descriptors can be found. For instance, detecting panoramic view shots and zooms can be useful for annotation purposes.

A motion-based approach has clear limitations. Similar motion patterns can be caused by different semantic concepts or events. The Bayesian framework used for classification allows us to overcome this limitation by embedding information from additional visual cues, like color and texture.

2.3 Other low-level features

2.3.1 Texture

Spatial textures have also been used to characterize semantic concepts. Picard and Minka use textural information in [64] to assist the user during the annotation process. In their system, the user provides a semantic label to one or several image regions and the label is automatically propagated to other visually similar regions of the image, in terms of their texture. The system knows several texture models and has the ability to choose the one that best explains the regions selected by the user, or even to create new explanations by combining models. The user can also provide negative examples to correct misclassifications and obtain more accurate explanations of semantic concepts. They show examples from their experiments using the following semantic labels: sky, grass, building, car, street and leaves. These concepts can thus be semantically characterized using textural information.

2.3.2 Orientation

Specific features extracted from textural information also characterize particular semantics of contents. For example, Gorkani and Picard use texture orientation in [31] to characterize "city/suburb" shots. Buildings, roads and other man-made structures found in city shots cause the presence of well defined orientations, particularly vertical and horizontal, while the orientations in nature scenes seem to be more random. Typical city and nature scenes are shown in fig. 2.7.

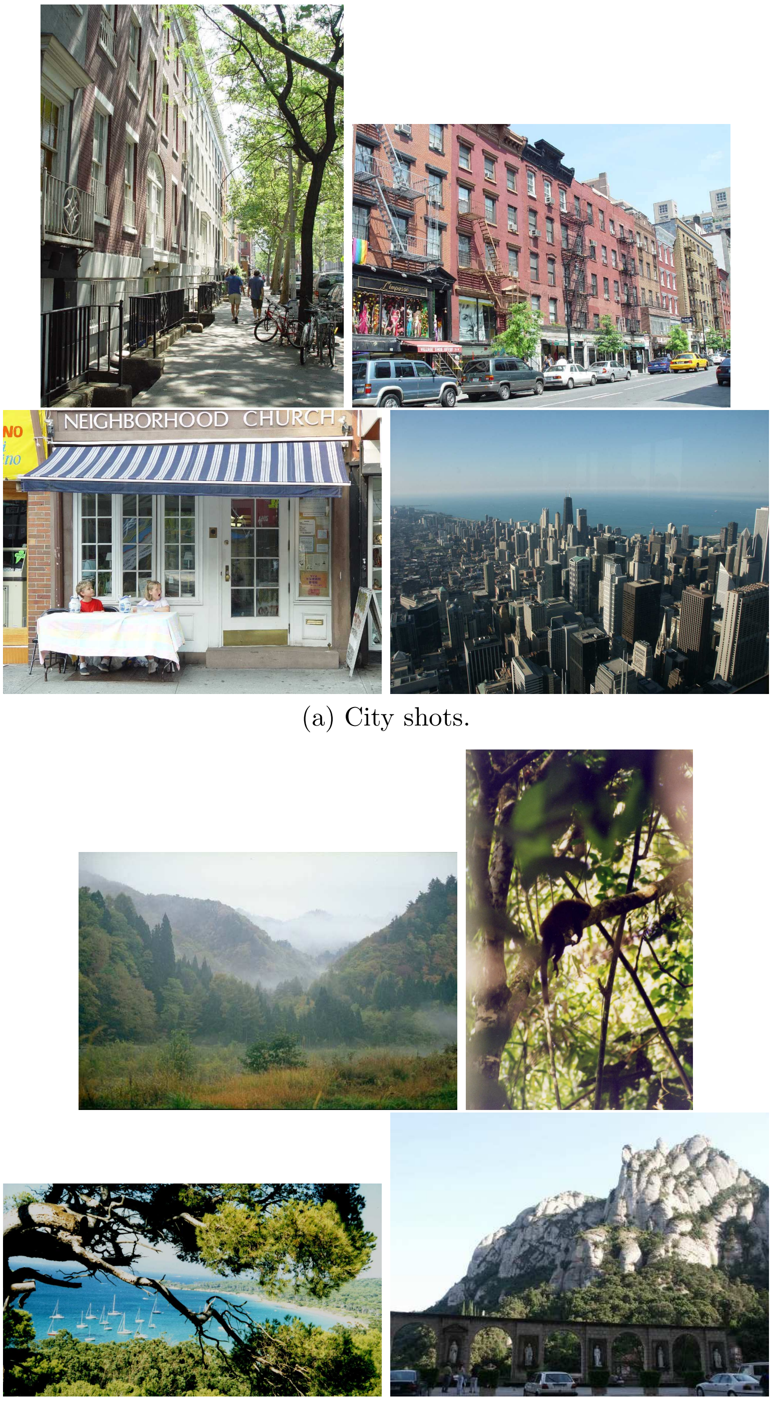
Following the same idea that city scenes can be characterized by the presence of man-made objects and structures, Vailaya et al. also deal with the classification of city vs. landscape images in [74]. Under this particular semantic classification problem, they evaluate the discriminative power of different low-level features, including color histogram, color coherence vector, DCT coefficients, edge direction histogram and edge direction coherence vector. They conclude that edge direction-based features have the most discriminative power. This conclusion could be expected a priori, as the main characteristic of man-made objects and structures is the presence of salient orientations, which are not represented by color-based features.

Orientation can also be considered from the semiotic point of view. Scenes shot with slanted slopes convey action, happiness and unreality, while horizontal and vertical slopes communicate calm. Dominant orientations are obtained using a modified Hough transform. In this case, the gradient magnitude is considered, so that lines with higher contrast will be enhanced in the transformed space of line parameters. This extension yields better results than the original Hough transform in terms of dominant orientations, as shown in fig. 2.8.

2.4 Combining multiple features

In the previous sections of this chapter, we have seen that different intermediate-level semantics can be attached to different low-level features. These relationships are summarized in table 2.3. It is reasonable to think that a combination of multiple low-level features will provide a better characterization of contents semantics than single-feature descriptions.

Several ways to combine information from multiple image features can be found in the literature. A first approach is to define a combined similarity measure, instead of a combined representation. This is the approach followed by Naphade et al. in [52]. They compute histograms for the following visual features: color, edge direction, motion magnitude and motion direction. They also consider audio features to obtain an audio-visual description of contents. Then, a "distortion" or distance measure is defined for each audio and visual feature. Finally, a weight is assigned to each distance



(a) City shots.

(b) Nature scenes.

Figure 2.7: Typical city (a) and nature (b) images. Man-made objects and structures present in city scenes show well defined orientations, while the orientations in nature scenes are more random.

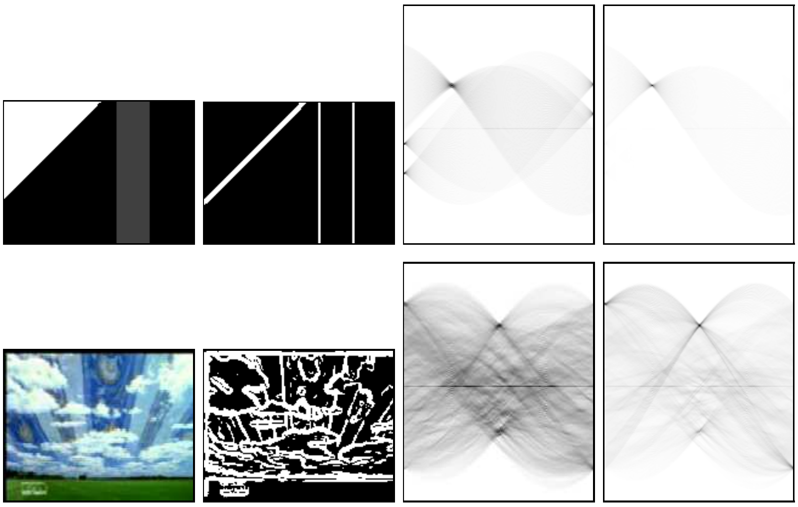


Figure 2.8: Examples of the Hough transform extended with gradient information. From left to right: original images, edge images (thresholded gradient), original Hough transform, and modified Hough transform.

Low-level feature	Intermediate-level semantics
Color	• Information about the appearance of an object, and thus its identity, is summarized in a color histogram. • When the object is the background, color provides information about location of the scene. • Semiotics associates colors with emotions conveyed to the viewer. Saturation, intensity and temperature also have emotional contents.
Motion	• Type of shot: close-up, medium shot, long shot, ... • Camera operation: pan, zoom, ... • Temporal motion textures can lead to the representation of complex concepts like “crowd”.
Spatial Texture	• Representation of concepts like sky, grass, leaves, building, car, street.
Orientation	• Complex concepts like “city vs. landscape”. • Semiotics associates slanted shots to dynamism and unrealistic scenes, and vertical/horizontal shots to calm.

Table 2.3: Summary of intermediate-level semantics that can be obtained from low-level visual features.

by the user in order to obtain the final "distortion", which is defined as:

$$D_O = \sum_{k=1}^{4} d_k(1 - w(k)) \tag{2.8}$$

where d_k and $w(k)$ are, respectively, the distance measure and the user-assigned weight for feature k. In this way, the user can define the significance of each feature in his/her query. For instance, if audio alone is used, clips that have explosion/gunshots/crashing followed by screams can be retrieved. With equal weights to audio and color, and the same query clip, clips with explosions and screams are returned. In this case, color imposes a more restrictive constraint and filters out clips with gunshots and crashings, given that they are visually different to explosions. An important advantage of this approach is that relevance feedback can be implemented by adjusting the weights according to the positive and negative examples provided by the user. Naphade et al. developed a weight updating strategy in [53].

Combining multiple features in the same similarity measure can be like putting peaches and melons in the same balance. For this reason, some authors try to use information from multiple features avoiding joining them. Ngo et al. implement in [55] a two-level hierarchical clustering of video shots, where color features are used at the top level, and motion at the bottom level. From table 2.3, their top level is clustering shots by their location or the objects in the scene, while at the bottom level they group by shot type, camera operation, or both. Vailaya et al. follow a similar approach in [73]. They define a semantic ontology for the hierarchical classification of vacation images, which is shown in fig. 2.9. A different classifier, based on different image features, is then used at each level of the hierarchy. The features involved in each classifier are summarized in table 2.4.

Szummer and Picard combine color and texture in [71] to face the problem of Indoor vs. Outdoor image classification in a similar fashion. Instead of combining features in the representation or in the similarity measure, they combine the output of multiple classifiers, one for each feature, using the majority function.

Information from multiple features can also be directly combined in the representation. Pass and Zabih propose the joint histogram in [61]. Each entry in a joint histogram contains the number of pixels in the image that are described by a particular combination of feature values. Therefore, a joint histogram is a multi-dimensional histogram, with one dimension (or more [1]) for each feature. The largest set of features considered in their work contains color, edge density, texturedness, gradient magnitude, and rank. This set of features yields a 7-dimensional joint histogram. The size of the data structure grows exponentially with the number of features. They report results with 4 to 5 quantization levels. However, if we had to consider 16 quantization levels per feature, the joint histogram would have 268,435,456 elements. The authors have considered this issue, and they report an average sparseness of 93% in the largest joint histograms. The problem is not only storage requirements, but also the time required for comparisons. On the other hand, there is not a way to know what features

[1] Color represented in the RGB color space yields 3 dimensions for one feature.

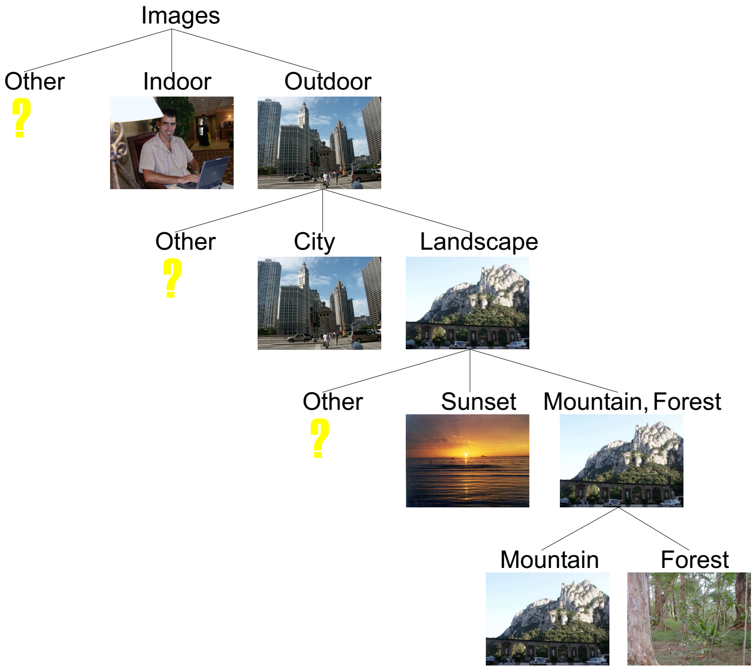


Figure 2.9: Hierarchical classification of vacation images by Vailaya et al. in [73].

Classification problem	Image features
Indoor vs. Outdoor	Spatial color and intensity distributions
City vs. Landscape	Distribution of edges
Sunset vs. Forest vs. Mountain	Global color distributions and saturation values

Table 2.4: Image features used in the different classification problems posed by Vailaya et al. in [73].

are more significant for the representation of some particular contents. In this way, irrelevant features could be discarded and removed from the representation, so that both storage size and comparison time would be reduced.

2.5 Summary

This chapter has analyzed the intermediate-level semantics that can be associated to different low-level image features. Color and motion are two main features that convey useful information for obtaining a higher level structure of the videos. Color histograms, and other extensions like the CCV, summarize the visual appearance of objects. When this object is the background of the scene, we obtain information about the location that, in most cases, is very relevant to group shots into story units. On the other hand, basic motion observations provide information about the type of shot (close-up, ...) and camera operation (zooming, ...), which can also be used as input to a high-level reasoning system for video structure analysis. Other image features like texture and orientation also provide relevant information for this purpose. The use of combined information from multiple low-level image features can lead to more robust and useful semantic descriptions of contents. However, the combination of features is difficult. Some methods require user intervention to specify the relevance of each feature, and naïve combinations turn into very demanding representations in terms of storage size and comparison time.