Compositional Object Recognition, Segmentation, and Tracking in Video

Björn Ommer and Joachim M. Buhmann[⋆]

Institute of Computational Science, ETH Zurich
8092 Zurich, Switzerland
{bjoern.ommer,jbuhmann}@inf.ethz.ch

Abstract. The complexity of visual representations is substantially limited by the compositional nature of our visual world which, therefore, renders learning structured object models feasible. During recognition, such structured models might however be disadvantageous, especially under the high computational demands of video. This contribution presents a compositional approach to video analysis that demonstrates the value of compositionality for both, learning of structured object models and recognition in near real-time. We unite category-level, multi-class object recognition, segmentation, and tracking in the same probabilistic graphical model. A model selection strategy is pursued to facilitate recognition and tracking of multiple objects that appear simultaneously in a video. Object models are learned from videos with heavy clutter and camera motion where only an overall category label for a training video is provided, but no hand-segmentation or localization of objects is required. For evaluation purposes a video categorization database is assembled and experiments convincingly demonstrate the suitability of the approach.

1 The Rational for Compositionality

Combined tracking, segmentation, and recognition of objects in videos is one of the long standing challenges of computer vision. When approaching real world scenarios with large intra-category variations, with weak supervision during training, and with real-time constraints during prediction, this problem becomes particularly difficult. By establishing a compositional representation, the complexity of object models can be reduced significantly and learning such models from limited training data becomes feasible. However, a structured representation might entail disadvantages during recognition, especially given the high computational demands of video. We present a compositional approach to video analysis that performs near real-time and demonstrates how the key concept of compositionality can actually be exploited for both, rendering learning tractable and making recognition computationally feasible.

Our compositional video analysis system unites category-level, multi-class object recognition, segmentation, and tracking in the same probabilistic graphical

[⋆] This work was supported in part by the Swiss national fund under contract no. 200021-107636.

A.L. Yuille et al. (Eds.): EMMCVPR 2007, LNCS 4679, pp. 318–333, 2007.

model. Moreover, this Bayesian network combines compositions together with object shape. Learning object models requires only a category label for the most prominent object in a complete video sequence, thereby even tolerating distracting clutter and other objects in the background. Category specific compositions of local features are automatically learned so that irrelevant image regions can be identified and discarded without supervision. As a result tedious hand-segmentations, object localizations, or initializations of a tracker become superfluous. Since there has been only very little work on category-level segmentation and recognition in video we have started assembling a video categorization database for evaluation purposes that consists of four object categories (bicycle, car, pedestrian, and streetcar). Videos have been recorded in their natural outdoor environment and show significant scale variation, large intra-category variability, camera panning, and background clutter.

Compositionality (e.g. [11]), which serves as a foundation for this contribution, is a general principle in cognition and can be especially observed in human vision [3]. Perception exhibits a strong tendency to represent complex entities by means of comparably few, simple, and widely usable parts together with relations between them. Rather than modeling an object directly based on a constellation of its parts (e.g. [9]), the compositional approach learns intermediate groupings of parts. As a consequence, compositions bridge the semantic gap between low level features and high level object recognition by modeling category-distinctive subregions of an object, which show small intra-category variations compared to the whole object. The robustness of compositions to image changes can be exploited for tracking and grouping them over consecutive video frames. This temporal grouping of compositions improves the compositional image representation and enhances object segmentation and recognition. To be able to simultaneously recognize multiple objects in a video, we have incorporated a model selection strategy that automatically estimates the correct model complexity based on a stability analysis.

2 Related Work

Category-level recognition, segmentation, and tracking of objects in videos is related to a number of subtasks. First, motion information can be exploited by selecting relevant features for tracking (e.g. [22]) and establishing correspondences between frames, e.g. using the method of Lucas and Kanade [15]. Second, most methods for recognition describe objects based on local descriptors such as *SIFT* features [14] or flow histograms [7], and template-based *appearance patches* (e.g. [1,13]). Combining local features in an object model can then proceed along several lines. A simple approach is to compute descriptors on a regular grid and concatenate all cells to obtain a joint model [7]. More complex representations of the spatial object structure are *constellation models* [9], hough voting strategies [13], many-to-many feature correspondences [8], image parsing graphs [24], and compositional models [18]. Viola and Jones have proposed a real-time recognition system for faces that is based on a cascade of classifiers [25]. Another

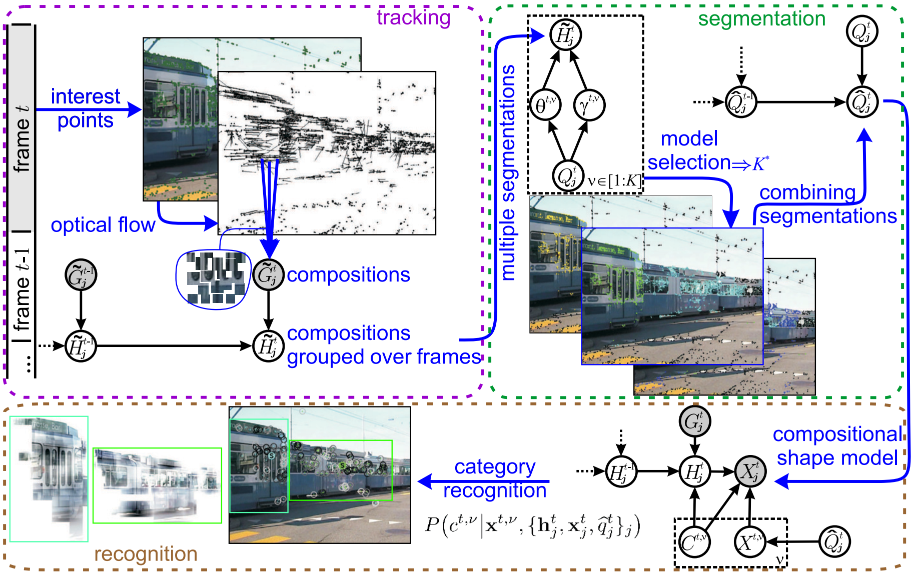


Fig. 1. Sketch of the processing pipeline for tracking, segmentation, and category-level object recognition in video

object class that many vision systems have been specifically developed for are pedestrians, e.g. [26,10]. Tracking algorithms have been studied for instance in [4] as well as [23], where the latter also presents a query-by-example approach to recognition that searches for regions which are similar to a user selected one. In contrast to tracking of a user specified region [5,2], Goldberger and Greenspann [12] propose a method for using segmentations of previous video frames to obtain a segmentation for the next.

3 Compositional Approach to Video Analysis

The following gives an overview of our compositional approach to video analysis (illustrated in Figure 1) before presenting the details in later sections. A novel video is analyzed sequentially in a frame-by-frame manner, while the underlying statistical model is propagating information over consecutive frames. Once a new frame is available, optical flow is estimated at interest points. These points and their motion pattern constitute the atomic parts of the composition system. Since interest points and their optical flow cannot be computed reliably, tracking individual points through a whole image sequence becomes error-prone. Therefore, we establish compositions of parts which are represented by probability distributions over their constituent parts. As a result, compositions are invariant with respect to individual missing parts and can be tracked reliably through a video by considering the optical flow distribution of all their constituents. The correspondence of compositions in consecutive frames is used to

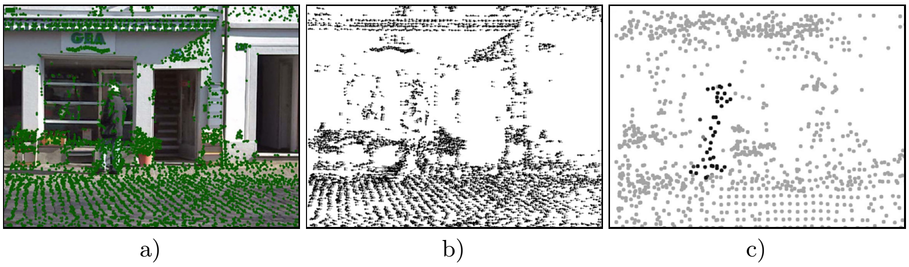


Fig. 2. a) Detected interest points. b) Estimated optical flow at interest points. c) Locations $\mathbf{x}_j^t$ of established compositions $\widetilde{\mathbf{h}}_j^t$. Brightness encodes the index of the closest codebook vector. See text for details.

group a composition over time. Subsequently, multiple segmentations are established. Therefore, compositions are clustered into different numbers of segments. To find an appropriate segmentation we incorporate a model selection strategy that analyzes the stability of the proposed segmentations over the preceding frames. The model with highest stability is then selected and combined with models from previous video frames to segment the current frame. Recognition of objects in the individual segments is then based on an extension of the *compositional shape model* from [18] which couples all compositions belonging to the same segment in a Bayesian network. The object category label is then obtained using probabilistic inference based on this model. In conclusion, tracking object constituents, segmenting objects from another, and recognizing the object category are all captured by the same statistical model, namely the graphical model illustrated in Figure 4. In this model, object representations of consecutive frames are linked together by a *Markov backbone* that connects segmentations of subsequent frames. Learning the underlying structured object model for a category proceeds in an unsupervised manner without requiring hand-segmentations or localization of objects in training videos.

3.1 Atomic Compositional Constituents

Based on the method of Shi and Tomasi [22] interest points are detected in every video frame, see Figure 2 a). Interest points from a preceding frame are then tracked into the next one using the Lucas-Kanade tracking algorithm [15]. This registration of points in consecutive frames yields an estimate of the optical flow $\mathbf{d}_i^t$ at interest point i in frame t, i.e. the displacement vector, see Figure 2 b). The interest points constitute the atomic parts of the composition system.

Codebook-Based Representation of Atomic Parts: Compositions can have different numbers of constituents and are, therefore, represented by a distribution over a codebook of atomic parts. Let $\mathbf{e}_i^t$ denote a feature vector that represents an atomic part i in frame t. The codebook that is used for representing compositions is then obtained by performing k-means clustering on all feature vectors

$\mathbf{e}_i^t$ from the training data. This vector quantization yields a common codebook of atomic parts for all object categories. To robustify the representation, each feature is described by a Gibbs distribution [27] over the codebook rather than by its nearest prototype: Let $d_\nu(\mathbf{e}_i^t)$ denote the squared euclidean distance of a measured feature $\mathbf{e}_i^t$ to a centroid $\mathbf{a}_\nu$. The local descriptor is then represented by the following distribution of its cluster assignment random variable F_i,

$$\begin{aligned} P(F_i = \nu | \mathbf{e}_i^t) &:= Z(\mathbf{e}_i^t)^{-1} \exp\left(-d_\nu(\mathbf{e}_i^t)\right) , \\ Z(\mathbf{e}_i^t) &:= \sum_\nu \exp\left(-d_\nu(\mathbf{e}_i^t)\right) . \end{aligned} \tag{1}$$

Local Descriptors: We use two different kinds of local features to represent local parts. The first type simply represents the optical flow at an interest point, whereas the second is based on *localized feature histograms* [17] of a small surrounding region. As optical flow has to be estimated for tracking in any case, this representation has the advantage that no extra feature detector needs to be computed at each interest point and, therefore, saves computation time. For each interest point i in frame t we use its optical flow $\mathbf{d}_i^t$, giving a 2-dimensional feature vector $\mathbf{e}_i^t = \mathbf{d}_i^t$.

The second local descriptor is formed by extracting quadratic image patches with a side length of 20 pixels at interest points. Each patch is divided up into four equally sized subpatches with locations fixed relative to the patch center. In each of these subwindows marginal histograms over edge orientation and edge strength are computed (allocating four bins to each of them). Furthermore, an eight bin color histogram over all subpatches is extracted. All these histograms are then combined in a common feature vector $\mathbf{e}_i^t$.

A separate codebook is established for both types of features (optical flow features are quantized with a 10 dimensional codebook, the localized feature histograms are represented by a 60 dimensional codebook). Tracking of compositions and object segmentation is then based on the optical flow alone. Only the final inference of the object category based on the compositions in a foreground segment uses the complex, second descriptor type.

3.2 Compositions of Parts

In the initial frame of a video ($t = 0$), a random subset of all detected interest points is selected. Each of these points is then grouped with the atomic parts in its local neighborhood (radius of 25 pixel) yielding compositions of atomic parts. A composition in frame t is then represented by a mixture distribution (with uniform mixture weights) over the densities (1) of its constituent parts (cf. [18]). Let $\Gamma_j^t = \{\mathbf{e}_1^t, \ldots, \mathbf{e}_m^t\}$ denote the grouping of parts represented by features $\mathbf{e}_1^t, \ldots, \mathbf{e}_m^t$. The multivariate random variable G_j^t does then represent the composition consisting of these parts. A realization $\mathbf{g}_j^t \in [0,1]^k$ of this random variable is a multivariate distribution over the k-dimensional codebook of atomic parts

$$\mathbf{g}_j^t \propto \sum_{i=1}^m \Big(P(F_i = 1 | \mathbf{e}_i^t), \ldots, P(F_i = k | \mathbf{e}_i^t) \Big)^T . \tag{2}$$

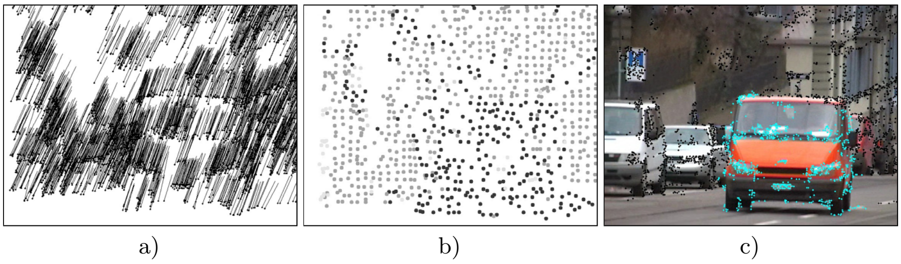


Fig. 3. Using compositions to obtain reliable segmentations despite strong camera panning. a) Estimated optical flow at interest points. b) Compositions $\widetilde{\mathbf{h}}_j^t$. c) Segmentation.

Finally, each of the k dimensions is independently standardized to zero mean and unit variance across the whole training set, giving *z-scores*. This mixture model has the favorable property of robustness with respect to variations in the individual parts. As we are having two types of features $\mathbf{e}_i^t$ we obtain two representations of a composition j: $\mathbf{g}_j^t$ is the representation based on localized feature histograms, whereas $\widetilde{\mathbf{g}}_j^t$ builds on optical flow.

Compositions are tracked throughout a video based on the average flow estimated at their constituent parts. Given the position $\mathbf{x}_j^t$ of a composition in frame t and the optical flow vectors of its parts $\mathbf{d}_i^t$, its predicted position in the next frame is

$$\mathbf{x}_j^{t+1} = \mathbf{x}_j^t + \frac{1}{m}\sum_{i=1}^{m} \mathbf{d}_i^t. \tag{3}$$

In the next frame $t+1$ the assignment of parts to compositions is updated since new interest points are computed. Therefore, all parts $\mathbf{e}_i^{t+1}$ in the local neighborhood of a composition $\mathbf{g}_j^{t+1}$ are assigned to this composition.

3.3 Temporal Grouping of Compositions

Whereas the preceding grouping was a spatial one (based on proximity) the following will present a grouping of compositions over time. By grouping compositions over consecutive frames, compositions of compositions can be formed that are more robust with respect to measurement errors in individual frames such as incorrect flow estimates. A temporal grouping of the j-th composition over consecutive frames yields the higher-order composition $\mathbf{h}_j^t$ which is represented by the distribution

$$\mathbf{h}_j^t \propto \begin{cases} \eta \mathbf{g}_j^t + (1-\eta)\mathbf{h}_j^{t-1}, & \text{if } t > 1\,, \\ \mathbf{g}_j^t, & \text{else}\,. \end{cases} \tag{4}$$

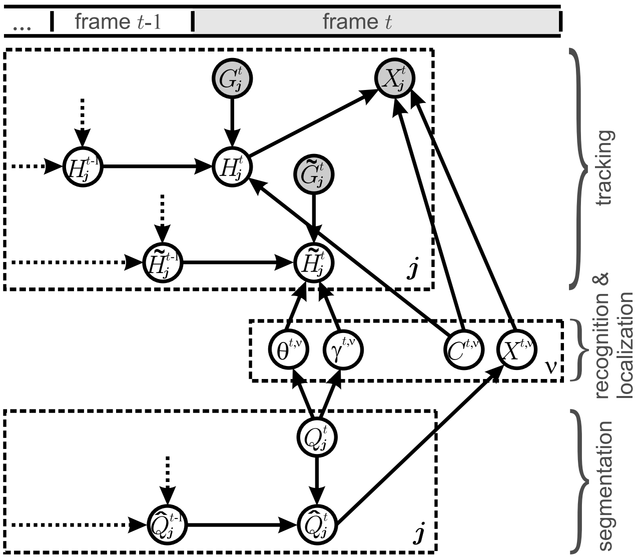


Fig. 4. Graphical model that unites category-level object recognition, segmentation, and tracking in the same statistical framework. Shaded notes denote evidence. The graph shows the dependencies between the three processes for frame t as well as the connection with the preceding frame. The involved random variables are the following: compositions represented with localized feature histograms, G_j^t, and with optical flow, $\widetilde{G}_j^t$. Temporal groupings of compositions: H_j^t and $\widetilde{H}_j^t$. Location of j-th composition: X_j^t. Assignment of compositions to segments: Q_j^t. Combining multiple segmentations over consecutive frames: $\widehat{Q}_j^t$. Segment priors: $\gamma^{t,\nu}$. Segment prototypes: $\boldsymbol{\theta}^{t,\nu}$. Classification of object in segment ν: $C^{t,\nu}$. Localization of the object: $X^{t,\nu}$.

The flow representation of compositions is computed according to the same recursion formula, i.e. $\widetilde{\mathbf{h}}_j^t \propto \eta\widetilde{\mathbf{g}}_j^t + (1-\eta)\widetilde{\mathbf{h}}_j^{t-1}$, and the mixture weight is chosen to be $\eta = 1/2$. The corresponding transition probability of the graphical model in Figure 4 is defined as

$$p(\mathbf{h}_j^t|\mathbf{g}_j^t, \mathbf{h}_j^{t-1}) := \mathbf{1}_{\{\mathbf{h}_j^t \propto \mathrm{Eq.(4)}\}} \in \{0,1\}. \qquad (5)$$

In Figure 2 c) the centers $\mathbf{x}_j^t$ of compositions $\widetilde{\mathbf{h}}_j^t$ are displayed. As described in Section 3.2, each composition is represented by a probability distribution over a codebook of atomic parts. The brightness of the circle at $\mathbf{x}_j^t$ encodes the index of the codebook vector that has received most probability mass. In Figure 3, strong camera panning results in unreliable optical flow estimates at interest points. Compositions, however, can compensate for this difficulty and establish the foundation for an accurate segmentation. In conclusion, the visualizations show that compositions are actually valuable for a subsequent segmentation of objects.

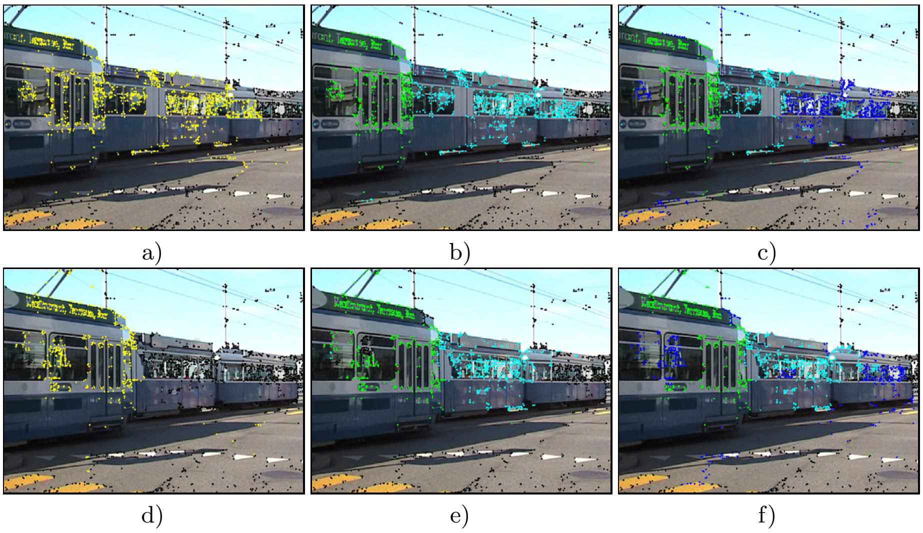

a) b) c)
d) e) f)

Fig. 5. Multiple segmentation hypotheses established for two frames. a) and d) show a 2 cluster solution. b) and e) 3 clusters. c) and f) 4 clusters. The segmentation with 3 clusters features the highest stability over the two frames and is, therefore, chosen by model selection.

3.4 Obtaining Multiple Segmentation Hypotheses

Subsequently, several initial hypotheses for the locations and shapes of objects that are present in a video frame are to be derived from the compositions. Since there is no prior information regarding the number of objects that are present in a scene, we have to address a difficult model selection problem. Therefore, several segmentations with varying numbers of segments are established. Model selection is then performed to retrieve the most reliable segmentation. Each segmentation partitions compositions in the optical flow feature space, $\widetilde{\mathbf{h}}^t_j$, into K segments using histogram clustering (e.g. see [19]): Compositions defined by (4) are represented as multivariate distributions over the k-dimensional part codebook, $\widetilde{\mathbf{h}}^t_j = (\widetilde{\mathbf{h}}^t_{j,1}, \ldots, \widetilde{\mathbf{h}}^t_{j,k}) \in [0,1]^k$ with $\sum_{l=1}^k \widetilde{\mathbf{h}}^t_{j,l} = 1$. The aim of clustering is then to represent $\widetilde{\mathbf{h}}^t_j$ by a mixture of K clusters $\boldsymbol{\theta}^{t,1}, \ldots, \boldsymbol{\theta}^{t,K} \in [0,1]^k$ with $\sum_{l=1}^k \boldsymbol{\theta}^{t,\nu}_l = 1$ and mixture weights or class priors $\gamma^{t,1}, \ldots, \gamma^{t,K} \in [0,1]$,

$$p\big(\widetilde{\mathbf{h}}^t_j \big| \boldsymbol{\theta}^{t,1}, \ldots, \boldsymbol{\theta}^{t,K}, \gamma^{t,1}, \ldots, \gamma^{t,K}\big) = \sum_{\nu=1}^{K} \gamma^{t,\nu}\, p\big(\widetilde{\mathbf{h}}^t_j \big| \boldsymbol{\theta}^{t,\nu}\big). \tag{6}$$

The individual mixture components are approximated by multinomial distributions, i.e. for large $N \in \mathbb{N}$ the distribution of $\widetilde{\mathbf{h}}^t_{j,l} \cdot N$ is multinomial with parameter $\boldsymbol{\theta}^{t,\nu}$. Transforming the definition of the multinomial distribution yields

$$
\begin{aligned}
p\left(\widetilde{\mathbf{h}}^t_j \middle| \boldsymbol{\theta}^{t,\nu}\right) &= \frac{N!}{\prod_l \lfloor \widetilde{\mathbf{h}}^t_{j,l} N \rceil!} \prod_l (\boldsymbol{\theta}^{t,\nu}_l)^{\widetilde{\mathbf{h}}^t_{j,l} N} & (7)\\
&= \frac{N!}{\prod_l \lfloor \widetilde{\mathbf{h}}^t_{j,l} N \rceil!} \frac{\exp\left(\sum_l \widetilde{\mathbf{h}}^t_{j,l} \log \widetilde{\mathbf{h}}^t_{j,l}\right)}{\exp\left(\sum_l \widetilde{\mathbf{h}}^t_{j,l} \log \widetilde{\mathbf{h}}^t_{j,l}\right)} \exp\Big\{ \sum_l \widetilde{\mathbf{h}}^t_{j,l} \log(\boldsymbol{\theta}^{t,\nu}_l)^N \Big\} & (8)\\
&= \frac{N!}{\prod_l \lfloor \widetilde{\mathbf{h}}^t_{j,l} N \rceil!} \prod_l (\widetilde{\mathbf{h}}^t_{j,l})^{\widetilde{\mathbf{h}}^t_{j,l}} \cdot \exp\Big\{ - \sum_l \widetilde{\mathbf{h}}^t_{j,l} \log \frac{\widetilde{\mathbf{h}}^t_{j,l}}{(\boldsymbol{\theta}^{t,\nu}_l)^N} \Big\} & (9)\\
&= \frac{N!}{\prod_l \lfloor \widetilde{\mathbf{h}}^t_{j,l} N \rceil!} \prod_l (\widetilde{\mathbf{h}}^t_{j,l})^{\widetilde{\mathbf{h}}^t_{j,l}} \cdot \exp\Big\{ -D_{KL}\left(\widetilde{\mathbf{h}}^t_j \| (\boldsymbol{\theta}^{t,\nu})^N\right) \Big\}. & (10)
\end{aligned}
$$

Here $D_{KL}(\cdot\|\cdot)$ denotes the Kullback Leibler distance between compositions and cluster prototypes while the prefactors are for normalization purposes.

The clusters $\boldsymbol{\theta}^{t,\nu}$ and the assignment $Q^t_j \in [1:K]$ of compositions to clusters, i.e. $P(Q^t_j = \nu) := \mathit{Prob}\{j\text{-th composition assigned to cluster } \nu\}$, are computed by iterating an *expectation-maximization algorithm* [16]. In the *expectation-step*, assignment probabilities of compositions to segments are computed conditioned on the current estimate of clusters,

$$
P\left(Q^t_j = \nu\right) := \frac{\gamma^{t,\nu}\, p\left(\widetilde{\mathbf{h}}^t_j | \boldsymbol{\theta}^{t,\nu}\right)}{\sum_\nu \gamma^{t,\nu}\, p\left(\widetilde{\mathbf{h}}^t_j | \boldsymbol{\theta}^{t,\nu}\right)}. \tag{11}
$$

In the *maximization-step*, class priors $\gamma^{t,\nu}$ and cluster prototypes $\boldsymbol{\theta}^{t,\nu}$ are updated conditioned on the assignment probabilities

$$
\gamma^{t,\nu} := \frac{\sum_j P(Q^t_j = \nu)}{\sum_{j,\nu'} P(Q^t_j = \nu')}, \quad \boldsymbol{\theta}^{t,\nu}_l := \frac{\sum_j P(Q^t_j = \nu)\, \widetilde{\mathbf{h}}^t_{j,l}}{\sum_j P(Q^t_j = \nu)}. \tag{12}
$$

After convergence of the EM-algorithm, the cluster assignment probabilities $P(Q^t_j = \nu)$ of compositions represent the segmentation of a video frame into K segments. Since background is surrounding objects, the segment that covers most of the frame border is labeled as background, $\nu = \text{BG}$. Figure 5 shows segmentations with 2, 3, and 4 segments for two video frames. Interest points in the different segments are displayed in distinct color (black is used for the background segment).

3.5 Model Selection to Identify Reliable Segmentation Hypotheses

As there is no prior information regarding the number of objects that are present in a scene we pursue a model selection strategy to estimate the number of object segments. Therefore, segmentations $Q^t_j(K)$ for different numbers K of segments are established in each frame (currently we use $K = 2, \ldots, 5$). Bipartite matching

[6] is performed to make the current segmentation comparable with the one of the previous frame, i.e. labels are permuted so that they fit best to the preceding segment labeling. We then combine multiple segmentations of consecutive video frames into a single, more robust one $\widehat{Q}_j^t(K)$, with

$$P\big(\widehat{Q}_j^t(K) = \nu \big| Q_j^t(K), \widehat{Q}_j^{t-1}(K)\big) \tag{13}$$
$$\propto \begin{cases} \eta P\big(Q_j^t(K) = \nu\big) + (1-\eta) P\big(\widehat{Q}_j^{t-1}(K) = \nu \big| Q_j^{t-1}(K), \widehat{Q}_j^{t-2}(K)\big), & \text{if } t > 1\,, \\ P\big(Q_j^t(K) = \nu\big), & \text{else}\,. \end{cases}$$

This dependency between segmentations of consecutive frames constitutes the Markov backbone that is represented at the bottom of the graphical model in Figure 4. It propagates segmentation hypotheses from previous frames into the current one.

An inappropriate model complexity is likely to yield unstable segmentations that change even when the input data varies only slightly. By observing the fluctuations of segmentations over multiple frames we can estimate their stability (cf. [20]) and select the most appropriate model complexity (see Figure 5 for an illustration). The stability $\zeta^t(K)$ of a K cluster segmentation is measured by the entropies $\mathcal{H}$ of the segment assignments

$$\zeta^t(K) := \sum_j \mathcal{H}\big(\widehat{Q}_j^t(K)\big) = -\sum_j \sum_{\nu=1}^{K} P\big(\widehat{Q}_j^t(K) = \nu\big) \log P\big(\widehat{Q}_j^t(K) = \nu\big). \tag{14}$$

The optimal number of segments is determined by selecting the $K^\star$ that minimizes this stability measure and we use the abbreviation $P(\widehat{q}_j^t) := P\big(\widehat{Q}_j^t(K^\star) = \widehat{q}_j^t\big)$.

The location of the ν-th segment center $\mathbf{x}^t(\nu)$ is estimated as the center of mass of all compositions assigned to this segment ($j \in \mathcal{A}_\nu^t$)

$$\mathbf{x}^{t,\nu} := \frac{1}{|\mathcal{A}_\nu^t|} \sum_{j \in \mathcal{A}_\nu^t} \mathbf{x}_j^t\,, \quad \mathcal{A}_\nu^t := \big\{ j : \nu = \operatorname*{argmax}_{\nu'} P\big(\widehat{Q}_j^t(K) = \nu'\big)\big\}. \tag{15}$$

3.6 Compositional Shape Model for Object Recognition

In every frame of a novel test video, the objects that are present in the individual segments have to be recognized. Therefore, all compositions $\mathbf{h}_j^t$, $j \in \mathcal{A}_\nu^t$ that are assigned to a segment ν are coupled in the graphical model shown in Figure 4. This statistical model is founded on the *compositional shape model* from [18]. The category $c^{t,\nu} \in \mathcal{L}$ of the object in segment ν can then be inferred from its posterior distribution

$$P\big(c^{t,\nu} \big| \mathbf{x}^{t,\nu}, \{\mathbf{h}_j^t, \mathbf{x}_j^t, \widehat{q}_j^t\}_j\big)$$
$$= \frac{p\left(\{\mathbf{h}_j^t, \mathbf{x}_j^t, \widehat{q}_j^t\}_{j \in \mathcal{A}_\nu^t}, \{\mathbf{h}_j^t, \mathbf{x}_j^t, \widehat{q}_j^t\}_{j \notin \mathcal{A}_\nu^t} \big| c^{t,\nu}, \mathbf{x}^{t,\nu}\right) P(c^{t,\nu}|\mathbf{x}^{t,\nu})}{p\left(\{\mathbf{h}_j^t, \mathbf{x}_j^t, \widehat{q}_j^t\}_j | \mathbf{x}^{t,\nu}\right)} \tag{16}$$

by applying Bayes' formula. Now the denominator can be skipped because it is independent of $c^{t,\nu}$. Furthermore, the category of an object should be independent of its absolute position in a frame and there should be no bias on any category, i.e. all classes are a priori equally likely. Therefore, $P(c^{t,\nu}|\mathbf{x}^{t,\nu})$ can be discarded as well.

$$\ldots \propto p\left(\{\mathbf{h}_j^t, \mathbf{x}_j^t, \widehat{q}_j^t\}_{j\in\mathcal{A}_\nu^t}, \{\mathbf{h}_j^t, \mathbf{x}_j^t, \widehat{q}_j^t\}_{j\notin\mathcal{A}_\nu^t} | c^{t,\nu}, \mathbf{x}^{t,\nu}\right). \tag{17}$$

Since the category of segment ν determines only compositions that have been assigned to this segment (i.e. $j \in \mathcal{A}_\nu^t$), all other compositions are independent of $c^{t,\nu}$ and can be skipped. Moreover, an assignment to segment ν implies $\widehat{q}_j^t = \nu$. Therefore $\widehat{q}_j^t$ can be dropped as well for $j \in \mathcal{A}_\nu^t$ and we obtain

$$\ldots \propto p\left(\{\mathbf{h}_j^t, \mathbf{x}_j^t\}_{j\in\mathcal{A}_\nu^t} | c^{t,\nu}, \mathbf{x}^{t,\nu}\right). \tag{18}$$

Compositions are conditionally independent, conditioned on the object model parameters $c^{t,\nu}$ and $\mathbf{x}^{t,\nu}$. Therefore, the likelihood factorizes and we can apply Bayes' formula again to obtain

$$\ldots \propto \prod_{j\in\mathcal{A}_\nu^t} \frac{P\left(c^{t,\nu}|\mathbf{x}^{t,\nu}, \mathbf{h}_j^t, \mathbf{x}_j^t\right) \cdot p\left(\mathbf{h}_j^t, \mathbf{x}_j^t | \mathbf{x}^{t,\nu}\right)}{P(c^{t,\nu}|\mathbf{x}^{t,\nu})}. \tag{19}$$

The factor $p\left(\mathbf{h}_j^t, \mathbf{x}_j^t|\mathbf{x}^{t,\nu}\right)$ does not depend on the object category and can be omitted. Moreover, the category of an object should be independent of its absolute position in a frame and there should be no bias on any category. Therefore, $P(c^{t,\nu}|\mathbf{x}^{t,\nu})$ can again be left out and we obtain

$$\ldots \propto \prod_{j\in\mathcal{A}_\nu^t} P\left(c^{t,\nu}|\mathbf{h}_j^t, S_j^{t,\nu} = \mathbf{x}^{t,\nu} - \mathbf{x}_j^t\right) \tag{20}$$

$$= \exp\Big[\sum_{j\in\mathcal{A}_\nu^t} \ln P\left(c^{t,\nu}|\mathbf{h}_j^t, S_j^{t,\nu} = \mathbf{x}^{t,\nu} - \mathbf{x}_j^t\right)\Big]. \tag{21}$$

Here the relative position of a composition with respect to the object center is represented by the shift $\mathbf{s}_j^{t,\nu} = \mathbf{x}^{t,\nu} - \mathbf{x}_j^t$. *Nonlinear kernel discriminant analysis* (NKDA) [21]) is used to estimate the distribution in (21). Therefore, probabilistic two-class kernel classifiers are trained on compositions extracted form the training data. These classifiers are coupled in a pairwise manner to solve the multi-class problem (see [21]). During recognition, an object can be recognized *efficiently* by applying the classifier to all compositions $\mathbf{h}_j^t$ and computing (21).

4 Evaluation

For still-image categorization large benchmark databases are available and the compositional approach has been shown (see [18]) to yield competitive performance compared to state-of-the-art methods in this setting. However, for the weakly supervised video analysis task that is pursued in this contribution there are, to the best of our knowledge, no comparable benchmark datasets available. Therefore, we have assembled a database for category-level object recognition in video consisting of 24 videos per category (categories car, bicycle, pedestrian, and streetcar). As can be seen from the examples in the figures, videos feature large intra-category variation (cf. Figure 6a and 7e), significant scale and viewpoint variation (e.g. Figure 7a, g), camera panning (cf. Figure 3), and background clutter. In the following, experiments are performed using 10-fold cross-validation. For each cross-validation step a random sample of 16 videos per category is drawn for training keeping the remainder for testing. Learning proceeds then on a randomly selected subset of 15 frames per video, while testing is performed on each frame. To avoid a bias towards categories with more test frames we average the retrieval rates for each category separately before averaging these scores over all frames. This evaluation approach has become the standard evaluation procedure in image categorization (e.g. see [18]).

4.1 Evaluating the Building Blocks of the Composition System

The following experiments evaluate the gain of the individual building blocks of the presented composition system for video analysis. In this first series of experiments only the most prominent object in a frame is to be detected. All key components are discarded in a first experiment before adding individual components in later experiments. The comparison of retrieval rates underlines the importance of each individual part of the compositional approach.

Baseline Performance of a Bag of Parts Approach: The compositional approach establishes an intermediate representation that is based on compositions of parts and the spatial structure of objects. In a first experiment this hidden representation layer is neglected to evaluate the gain of compositionality. A frame is then represented based on all detected localized histogram features $\mathbf{e}_i^t$ by a bag of parts $\mathbf{b}^t$ (cf. Section 3.1),

$$\mathbf{b}^t \propto \sum_i \Big(P(F_i = 1|\mathbf{e}_i^t), \ldots, P(F_i = k|\mathbf{e}_i^t) \Big)^T . \tag{22}$$

To categorize a frame, the category with highest posterior $P(c^t|\mathbf{b}^t)$ is selected. The posterior is again learned from the training data using NKDA. This approach yields a retrieval rate of $\mathbf{53.1} \pm \mathbf{5.5}\%$.

Compositional Segmentation and Recognition w/o Shape Model: This experiment shows the benefit of combining segmentation with recognition in a compositional model. Therefore, compositions are established as described in Section 3.2 and Section 3.3. The prominent object in a video frame is then segmented from background clutter by establishing a 2-class segmentation as described in Section 3.4. Since only a single segmentation hypothesis is established no model selection is required. All compositions that are not assigned to the background, $\nu \neq \mathrm{BG}$, are then taken into account to recognize the most prominent object. Therefore, these compositions are combined using a bag of compositions descriptor $\widetilde{\mathbf{b}}^t \propto \sum_{j \in \mathcal{A}^t_{\nu \neq \mathrm{BG}}} \mathbf{h}^t_j$. Frames are then categorized without the compositional shape model by simply selecting the category with highest posterior $P(c^t|\widetilde{\mathbf{b}}^t)$. The combination of foreground segmentation with recognition based on compositions improves the retrieval rate to $\mathbf{64.5 \pm 5.5}\%$.

Segmentation, Recognition, and Compositional Shape Model: In contrast to the previous experiment we now use the full compositional shape model of Section 3.6 to recognize the foreground object. As a result, the retrieval rate is further increased to $\mathbf{74.3 \pm 4.3}\%$. The category confusion table is presented in Table 1. Another setting is to categorize video sequences as a whole and not individual frames. For this task the category hypothesis that is most consistent with all frames of a video is chosen. Here the compositional model achieves a retrieval rate of $\mathbf{87.4 \pm 5.8}\%$. Obviously, this is an easier setting since information from an ensemble of frames can be used simultaneously.

By agglomerating atomic parts of limited reliability in compositions that can be tracked reliably, information has been condensed and the robustness of object representations has been improved. The underlying statistical inference problem can then be solved efficiently. As a result the compositional model segments, tracks, and recognizes objects in videos of full PAL resolution (768× 576 pixel) at the order of 1 fps on an ordinary desktop computer.

Table 1. Category confusion table (percentages) for the complete composition system

True classes →	bicycle	car	pedestrian	streetcar
bicycle	**70.1**	5.5	15.1	4.0
car	10.0	**87.6**	16.1	12.4
pedestrian	15.9	2.5	**61.4**	5.5
streetcar	4.0	4.4	7.4	**78.2**

4.2 Multi-object Recognition

In the following experiment multiple objects that appear simultaneously in a video are to be recognized. Therefore, the model selection strategy of Section 3.5 is applied to find the correct number of objects. A frame is then correctly classified if all the present objects have been found. Missing an object or detecting an object that is not present counts as an error. Given this significantly harder task our full compositional model classifies $68.1 \pm 4.9\%$ of all frames correctly.

4.3 Analyzing the Relevance of Compositions

To analyze what individual compositions contribute to object recognition the category posterior

$$P\big(c^{t,\nu}\big|\mathbf{x}^{t,\nu}, \mathbf{h}^t_j, \mathbf{x}^t_j, \widehat{q}^t_j\big)\big|_{c^{t,\nu} = \text{ True Category}} \tag{23}$$

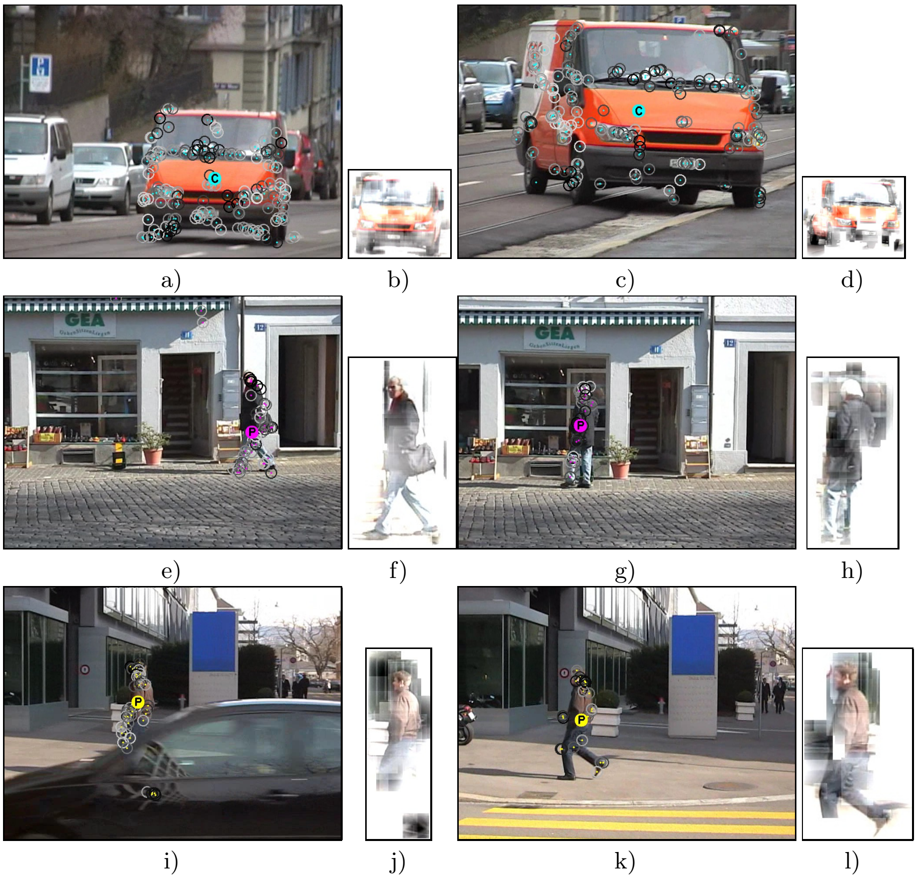
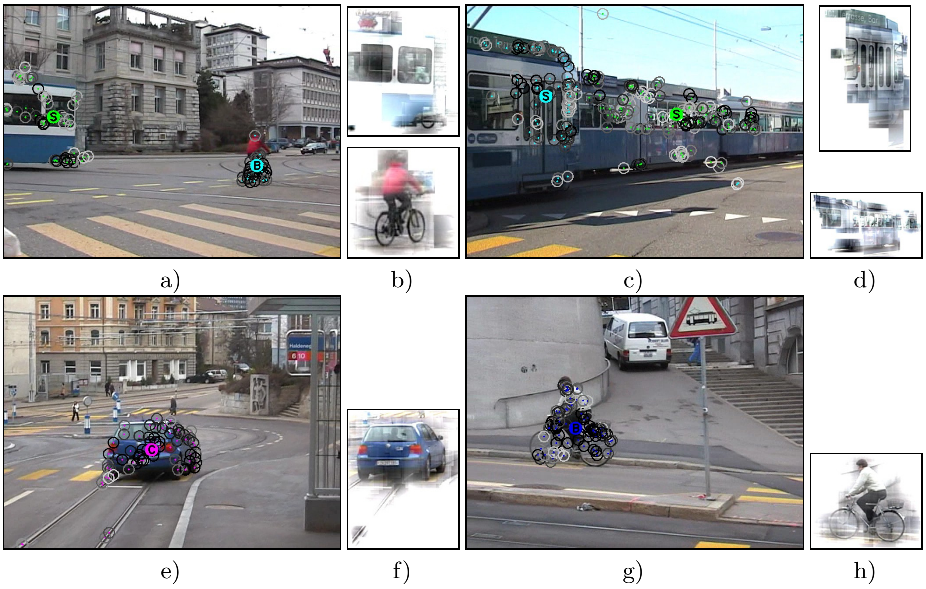


Fig. 6. Visualizing the contributions of compositions $\mathbf{h}_j^t$ to object recognition. Dark circles correspond to compositions with high posterior from (23). The gap between a) and c) and between e) and g) is both times 60 frames. i) and k) have a gap of 10 frames. Class labels are placed at the location of the segment center $\mathbf{x}^{t,\nu}$ (c:car, p:pedestrian).

is evaluated for each composition. In Figure 6 and 7 the category posterior is then encoded i) in the darkness of a circle around the composition center and ii) in the opaqueness of the underlying image region, i.e. alpha blending is used for visualization. Moreover, Figure 6 shows the propagation of an object segmentation over several frames.

5 Discussion

In this contribution we have presented a compositional approach that combines category-level object recognition, segmentation, and tracking in the same graphical model without user supervision. The compositional representation of objects



a) b) c) d)

e) f) g) h)

Fig. 7. Visualization of compositions. See Figure 6 for details.

is automatically learned during training. A model selection strategy has been pursued to handle multiple, simultaneously appearing objects. By agglomerating ensembles of low-level features with limited reliability, compositions with increased robustness have been established. As a result an intermediate object representation has been formed that condenses object information in informative compositions. Recognition has been formulated as an efficiently solvable, statistical inference problem in the underlying Bayesian network. Therefore, compositionality not only improves the learning of object models but also enhances recognition performance so that near real-time video analysis becomes feasible.

References

1. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. PAMI, 26(11) (2004)
2. Avidan, S.: Ensemble tracking. In: CVPR (2005)
3. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review, 94(2) (1987)
4. Brostow, G.J., Cipolla, R.: Unsupervised bayesian detection of independent motion in crowds. In: CVPR (2006)
5. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. PAMI, 25(5) (2003)
6. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)

7. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV (2006)
8. Demirci, F., Shokoufandeh, A., Keselman, Y., Bretzner, L., Dickinson, S.: Object recognition as many-to-many feature matching. IJCV, 69(2) (2006)
9. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)
10. Gavrila, D.M., Giebel, J., Munder, S.: Vision-based pedestrian detection: The protector+ system. In: IEEE Intelligent Vehicles Symposium (2004)
11. Geman, S., Potter, D.F., Chi, Z.: Composition Systems. Quarterly of Applied Mathematics, 60 (2002)
12. Goldberger, J., Greenspann, H.: Context-based segmentation of image sequences. PAMI, 28(3) (2006)
13. Leibe, B., Schiele, B.: Scale-invariant object categorization using a scale-adaptive mean-shift search. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) Pattern Recognition. LNCS, vol. 3175, Springer, Heidelberg (2004)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV, 60(2) (2004)
15. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI (1981)
16. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. John Wiley & Sons, Chichester (1997)
17. Ommer, B., Buhmann, J.M.: Object categorization by compositional graphical models. In: Rangarajan, A., Vemuri, B., Yuille, A.L. (eds.) EMMCVPR 2005. LNCS, vol. 3757, Springer, Heidelberg (2005)
18. Ommer, B., Buhmann, J.M.: Learning compositional categorization models. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, Springer, Heidelberg (2006)
19. Puzicha, J., Hofmann, T., Buhmann, J.M.: Histogram clustering for unsupervised segmentation and image retrieval. Pattern Recognition Letters, 20 (1999)
20. Roth, V., Lange, T.: Adaptive feature selection in image segmentation. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) Pattern Recognition. LNCS, vol. 3175, Springer, Heidelberg (2004)
21. Roth, V., Tsuda, K.: Pairwise coupling for machine recognition of hand-printed japanese characters. In: CVPR (2001)
22. Shi, J., Tomasi, C.: Good features to track. In: CVPR (1994)
23. Sivic, J., Schaffalitzky, F., Zisserman, A.: Object level grouping for video shots. IJCV, 67(2) (2006)
24. Tu, Z.W., Chen, X.R., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection and recognition. IJCV, 63(2) (2005)
25. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
26. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: ICCV (2003)
27. Winkler, G.: Image Analysis, Random Fields and Markov Chain Monte Carlo Methods—A Mathematical Introduction, 2nd edn. Springer, Heidelberg (2003)