Machine Vision and Applications (2011) 22:99–115
DOI 10.1007/s00138-009-0217-8

ORIGINAL PAPER

Multimedia translation for linking visual data to semantics in videos

Pınar Duygulu · Muhammet Baştan

Received: 25 July 2008 / Revised: 1 March 2009 / Accepted: 28 July 2009 / Published online: 30 September 2009

Abstract The semantic gap problem, which can be referred to as the disconnection between low-level multimedia data and high-level semantics, is an important obstacle to build real-world multimedia systems. The recently developed methods that can use large volumes of loosely labeled data to provide solutions for automatic image annotation stand as promising approaches toward solving this problem. In this paper, we are interested in how some of these methods can be applied to semantic gap problems that appear in other application domains beyond image annotation. Specifically, we introduce new problems that appear in videos, such as the linking of keyframes with speech transcript text and the linking of faces with names. In a common framework, we formulate these problems as the problem of finding missing correspondences between visual and semantic data and apply the multimedia translation method. We evaluate the performance of the multimedia translation method on these problems and compare its performance against other auto-annotation and classifier-based methods. The experiments, carried out on over 300 h of news videos from TRECVid 2004 and TRECVid 2006 corpora, show that the multimedia translation method provides a performance that is comparable to the other auto-annotation methods and superior performance compared to other classifier-based methods.

P. Duygulu (✉) · M. Baştan
Department of Computer Engineering,
Bilkent University, Ankara, Turkey
e-mail: duygulu@cs.bilkent.edu.tr

M. Baştan
e-mail: bastan@cs.bilkent.edu.tr

1 Introduction

Rapidly growing quantities of multimedia data, such as digital image and video archives, have created a demand for methods that can perform effective and efficient indexing, retrieval, and analysis of such data [1,2]. However, there is a gap between what users need, and what current methods can provide: users often wish to retrieve data based on their semantic content, whereas existing methods can work only on rather low-level visual representation of multimedia data. This gap, described as the semantic gap, refers to "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [1], and has become one of the central problems in multimedia. In [3], it is argued that the semantic gap consists of a hierarchy of gaps, and that it can be characterized by mainly two different types of gaps: namely, the gap between descriptors and labels, and the gap between labels and full semantics. We use the term semantic gap to denote the gap between descriptors and labels.

Current studies in image and video retrieval usually confine themselves to either text [4] or low-level features [1,5,6], or to simple combinations of text and low-level features [7–10]. The bridging of the semantic gap by linking the low-level descriptors, extracted from the visual data with the symbolic labels obtained from semantic data, is of utmost importance to improve the capability of these systems.

In most of the current systems, semantics is provided through manual labeling, and classifiers are then trained in a supervised way to link the low-level features with labels. The literature is broad and includes many different methods for detecting and recognizing specific objects (e.g., cars or pedestrians), faces, and scenes [11–16]. However, most of these methods suffer from two major drawbacks.

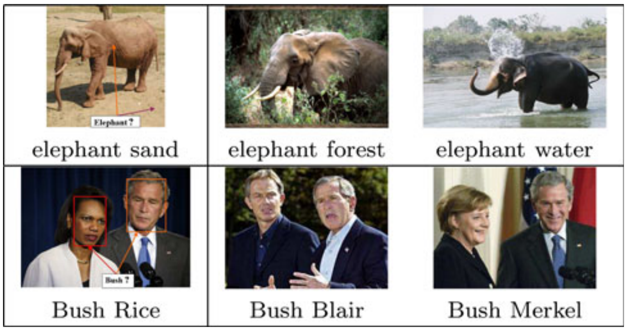


Fig. 1 Example images with annotation keywords. Although the images include the referred objects or faces, the exact locations in the images are unknown

First, in order to train a supervised system for object classes or for specific objects, labels need to be manually provided. This process requires a lot of human effort and can generate only subjective and error-prone semantic information. Therefore, most studies use small and controlled data sets with only few specific categories. Moreover, an image in such data sets is usually uncluttered and contains a single, centered instance of a specific category (e.g., an airplane, a face, a building in the middle and covering a large portion of the image). The number of categories has been increased in some of the data sets including Caltech data sets [17–19], PASCAL data set [20,21], and LabelMe data set [22,23], but the issue of uncluttered content usually remains.

A second and more important problem is that, due to the variety of semantic concepts, separate models must be built for each concept. However, this is difficult, if not impossible, when the number of categories to be learned is large [24]. Therefore, most of the current methods can work only in limited domains and only for a few distinct object categories or faces in controlled environments, and hence are not suitable for building large-scale systems.

Recently, it became a challenge to discover the links between visual and semantic information from the large annotated image collections. The labels in such data sets can be manually entered by a few people (such as stock photographs or museum images [25]); can be generated through community tagging (such as Flickr [26]) or through games (such as ESP game [27,28]); or can be made available without any additional effort due to the nature of the data (news photographs with captions [29]). In any case, it is easier to generate these data sets compared to the data sets discussed previously, and therefore they are available in larger volumes.

However, these data sets provide only loosely labeled data, such as the keyword descriptions for the entire image, but do not provide localized semantic data, such as the labels for the segmented regions of an image. Although we know "what the image is about" we do not know "where the object is specifically".

Let us consider the examples in Fig. 1. Given a single image and no prior knowledge, the correspondences between visual and semantic information are ambiguous. Semantic information provided through keywords asserts that the images have the referred objects or faces. However, it is not known which part of the visual information is linked to that semantic information. For example, in the top leftmost image we know that there is an elephant in the image but we do not know where it is. Similarly in the bottom leftmost image, we know that it contains the face of Bush, but we do not know which one it is.

Moreover, since the number of categories in such data sets is specified by the vocabulary of keywords used in labeling, the number of categories is generally in the order of hundreds. The labels are usually incomplete, subjective or error-prone [30]. There are some keywords which do not correspond to any visual information, such as the abstract words, or may correspond to a more general description which cannot be matched with a single piece of visual information, such as the city/country names [31]. Similarly, there are some regions/areas in the image which cannot be described with any keyword.

Due to all these problems, this type of data cannot be directly used within a supervised classification scheme to learn specific categories as required by most of the traditional systems.

In recent studies, as an alternative to supervised methods, machine learning techniques mostly adapted from text retrieval literature, have been used to link visual data with semantics in loosely labeled data sets. These studies handle the aforementioned problems using concurrent occurrences of visual and semantic information, and discarding the noisy information automatically using large volumes of data itself.

While such methods are mostly utilized for retrieval and automatic image annotation problems, we argue that, they are also relevant to the solution of a more general semantic gap problem. This relevance is also stated by Hare et al. [3] as "auto-annotation attempts to bridge the gap between descriptors and symbolic labels by learning which combinations of descriptors represent objects, and what the labels of the objects should be".

In order to understand this idea, consider the examples in Fig. 1 again. If we have other images where the elephant is in forest, or in water; and where Bush is together with Blair, or Merkel, then we can capture the link between the brown region and the elephant; the link between the face and the name of Bush; or in general the link between an object and its label as an important solution to recognition in the large scale.

In the following, we first review the studies for automatic image annotation by grouping them into three categories, and discuss their applicability to the semantic gap problem. Inspired by the challenges in large video collections, we then present new problems where it is also required to link visual data to semantics going beyond image annotation. We then show that various problems can be solved through a common framework.

2 Review and discussion of related studies in automatic image annotation

The goal of automatic image annotation is to predict the descriptive keywords for an image using the knowledge learned from other annotated images. Recently, there have been many attempts to solve the automatic image annotation problem. These studies can be grouped into three categories based on how they learn the links between visual and semantic data.

The methods in the first category view image annotation as a classification problem and classify the entire image or its parts into annotation keywords. In [32], labeled images are used as bags of examples and multiple-instance learning is proposed to classify images. Separate classifiers are built for each concept; an image is positive if a given concept (e.g., a waterfall) is present somewhere in it and negative if the concept is absent. Using a similar formulation, in [33], a Maximum Entropy based approach is proposed and multiple binary classifiers are built to annotate the images. In [34], the problem is formulated as M-ary classification in which each of the semantic concepts of interest defines an image class and the classes compete directly at the time of annotation. In [35], image concepts are modeled by 2-D multi-resolution Hidden Markov Models; an image is labeled with the concepts that best fit the content. In [16], a visual scene modeling and classification approach using a combination of text modeling methods and local invariant features is presented. It uses *bag-of-visterms* representation in a multi-class classification scheme and learns N concepts by training N SVMs (one-against-all). In testing phase, each test image is assigned to the class of the SVM that has the highest output of its decision function.

The methods in the second category model the joint distribution of words and image regions in the annotated image collections with the assumption that either the image regions and words are independent [36–38], or there is a conditional relationship provided by a hidden aspect variable [24,39,40]. The image regions are either represented by continuous descriptors or in the form of blobs which refer to the labels of clusters that are obtained by vector quantization of region descriptors. In [36], a Cross-media Relevance Model is introduced and a training set of annotated images is used to estimate the joint probability of observing a word with a set of blobs in the same image. In subsequent studies, discrete blob representation is replaced with the direct modeling of continuous descriptors of the regions, and two new models are proposed: the Continuous Relevance Model [37] and the Multiple Bernoulli Relevance Model [38]. The model proposed in [39,41] is a generative hierarchical aspect model inspired by the model proposed for text [42]; it combines the aspect model with a soft clustering model. Images and corresponding words are generated by nodes arranged in a tree structure. Image regions represented by continuous descriptors are modeled using a Gaussian distribution, and words are modeled using a multinomial distribution. In [40], the Corr-LDA (Correspondence Latent Dirichlet Allocation) model is proposed; this model finds the conditional relationships between latent variable representations of sets of image regions and sets of words. The model generates first the region descriptions and then the caption words. In [43], the probabilistic latent space models are modified to give higher importance to words. First the definition of latent space is constrained by focusing on words, then visual variations are learned conditioned on the space learned from words.

The methods in the third category learn the direct correspondences between image regions and words [44–47]. The first model proposed by Mori et al. [44] learns the joint distribution of blobs and words using the co-occurrence statistics. In [45], learning the correspondences between blobs and words is tackled as a translation problem. A probability table linking blobs and words is learned using Statistical Machine Translation techniques [48]. In [46], correlations between blobs and words are discovered based on co-occurrence counts and also on the cosine similarity of occurrence patterns—the documents including those items. The results are observed to be better if words and blobs are weighted in inverse proportion to their occurrence and Singular Value Decomposition is applied to suppress the noise. In [47], the spatial context is considered and the probability of a blob being aligned with a particular word is estimated depending on the word assignments of its neighboring blobs using Markov Random Fields.

Here, we discuss the three categories according to their applicability to the semantic gap problem. The classifier-based approaches in the first category require the learning of one classifier for each word or for a set of words, and therefore are not scalable. Although higher performances are reported with the models in the second category for annotation and retrieval purposes, these methods have two drawbacks. First, they do not learn exact mappings between regions and words; hence it is difficult to generalize these approaches for a larger domain of semantic gap problems in which labels should be linked with regions, such as for the recognition of objects and faces. Second, the links between the regions and keywords are learned through an image. Although context information is shown to be helpful in many recognition tasks, relying on context has some drawbacks. For instance, if all we learned is that tigers are on grass then we cannot recognize the tigers at the circus, or if all we learned is that person A is with person B, then we cannot recognize him when he is with C. Based on these insights, we argue that approaches in the last category better address the semantic gap problem.

When considered together, there is a common characteristic among all the three categories. They are all limited with the semantic concepts that can be described visually and thus may not fully satisfy the users which look for abstract concepts [31].

Prior to all the learning methods mentioned above, visual data is required to be processed. Here, we briefly summarize some of the methods used for this purpose.

For segmentation into regions, Normalized Cuts image segmentation algorithm [49] is used in the data set generated by [45]. In [50], several other segmentation algorithms are evaluated and it is shown that the success of image segmentation algorithms has an effect on the image annotation performance. But it is only a slight difference, and in [51], it is claimed that using a fixed grid partitioning is usually sufficient.

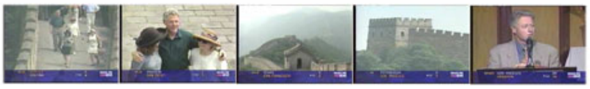
Feature selection is another criterion slightly affecting the performance [52]. While color and texture features extracted from regions have been heavily experimented (see [50] for a comparison), features extracted from local descriptors have also been used [53].

Selection of the number of clusters for the methods that use quantized representations is also another decision affecting the results in the form of polysemy (same blob may correspond to different visual information), or synonymy (different blobs may correspond to the same visual information). Several values are empirically evaluated and usually clusters in the order of hundreds are used.

3 Motivation

Video collections are available in huge volumes creating a demand for efficient access to the content. Toward this end, TRECVid video retrieval evaluation organized by NIST [54,55] is an important effort in providing benchmark data sets. As also addressed with the high-level feature task of TRECVid, automatic labeling of video content is an important challenge [56]. The requirement for scalable methods that can learn large number of concepts with minimal effort as an alternative to supervised systems suggests the application of image annotation techniques to video data sets. However, while the Corel data set consisting of stock photographs with a few annotation keywords—provided in [39,45]—was extensively studied in image annotation literature as being a more difficult and noisier one, TRECVid data set is studied in a similar setting by only a few groups (e.g., [38,57,58]).

In TRECVid data set, keyframes for a small number of videos are manually annotated with a collaborative effort. The common approach in the image annotation studies working on TRECVid data is to treat the video keyframes as if they are images and then to directly apply their image annotation methods to annotate keyframes. However, such methods are limited with the number of annotation words provided by the lexicon in the training set.



... (1)earlier today mr. **clinton** toured some of china's most ancient (2) sites including the **great wall** and the forbidden city he started his (3) day by attending a sunday worship service and one of china's non denominational churches from the public mr. **clinton** (4) praised the growth of religious freedom in china we celebrate (5) with you the growth of the practice of our faith in china e rejoice to hear rivers or cite the numbers of churches and other places of worship where people are practicing their faith today...

Fig. 2 A video story from broadcast news about Clinton's visit to China. The speech transcripts temporally aligned with the video keyframes do not always match with the visual appearances. For example, while the names of *Clinton* and *Great Wall* are mentioned in the speech transcript of shots (1,3) and (2), the shots corresponding to visual appearances are (2,5) and (1,3,4), respectively

An alternative approach is to use the speech transcript text, which, although noisy, is available for the entire set of videos. The speech transcript text is usually aligned with the closed caption text [59]. This process, which is usually referred to as video alignment, aligns the text with video frames only temporally but not related to the visual content. Therefore, even after video alignment the links between visual and semantic information are unknown. We refer to this problem as video association problem.

For a better understanding of the problem, let us consider an example. Usually, in broadcast news, an anchorperson mentions an event when he/she is introducing the story, but the event visually appears later in the story, probably with an aligned text which does not include any related word to the event. For instance, in the story about Clinton's visit to China shown in Fig. 2, the names of Clinton and Great Wall are mentioned in the speech transcripts, but it is not known which shot includes the corresponding visual appearances. Since the time difference between the visual and textual descriptions can be large, or the order can be different, using a set of shots in a neighborhood may still not solve such a problem.

Another important challenge, especially for news videos, is to handle queries related to people that lie at the core of most stories. Searching for names in the speech transcript text is likely to produce incorrect results due to the incorrect alignments mentioned earlier. Moreover, the shots associated with the text may include no people at all; many other people besides the target; or another person, particularly the anchorperson or reporter. On the other hand, the recognition of faces is a long-standing and difficult problem [60,61]. It is still a challenge to recognize faces in different poses, environments, or illuminations and without affected by occlusion, clutter, aging, clothing or make-up.

The problem of labeling large number of faces in realistic environments as in TRECVid videos, which we refer to as face naming, requires alternative solutions, and can also be tackled by associating faces and names in large collections. As an attempt in this direction, in [62], face and name information is integrated to improve the performance of person queries by modeling the timing between the naming and appearance of people in news videos. In some other studies faces are grouped given the names [63,64]. Initially, the faces appearing in a news photograph are assumed to be linked to the names in the caption, and then different methods are used to correct these links. The most relevant solution in linking faces with names is the Name-It system [65], which aims to associate faces and names in news videos using co-occurrence statistics.

3.1 Discussion on the challenges in videos

Here, we summarize the challenges through the perspective of image annotation approaches.

- *Providing alternatives to manual annotations.* As discussed earlier, the scalable systems should not depend on manual annotations which are difficult to obtain. The alternative is to use the available speech transcripts as semantics to be linked to visual information [58,66].
- *Going from images to video stories.* In a video, not a single keyframe, but a sequence of keyframes are associated with the speech transcript text. A story is an ideal unit to associate visual and semantic information in a video. However, this provides rough association and it is thus required to learn the direct links between the visual units in a story and the words in speech transcripts. This problem suggests the use of image annotation techniques to be applied on video stories by widening the perspective.
- *Generalization of the visual information.* The labeling of faces is not different from the labeling of objects if the problem is turned into the problem of finding the links between visual information (faces in this case) and semantic information (names in this case). The only difference is in the representation of faces. This suggests the utilization of image annotation studies not only for labeling of objects or scenes but also for labeling of faces in a similar formulation.

4 Proposed approach

We argue that the new problems together with image annotation can be formulated commonly as the problem of missing correspondences between visual and semantic data. In all cases, the semantics is associated with visual information only loosely within a given image or video story, but the correspondences are missing. However, the concurrent appearance of the same visual and semantic information in other images or videos provides hints about the correspondences.

In the following, we describe a framework to formulate a wide range of problems as the problem of missing correspondences between low-level visual data and semantics, and discuss how the gap between them can be bridged based on the idea of capturing the co-occurrences. Building on [39,45], we then utilize the multimedia translation method to link visual and semantic data.

4.1 Formulation of the problem as missing correspondences

Consider a multimedia collection $M = \{m^1, \ldots, m^N\}$, where each $m^n = (\mathbf{v}^n, \mathbf{s}^n)$ is an associated pair of *visual content* $\mathbf{v}^n$ and *semantic content* $\mathbf{s}^n$, such as an image annotated with keywords or a video with associated speech transcript text (see Fig. 3).

Both type of data may need some processing to obtain informative descriptions. The visual content may be pro-

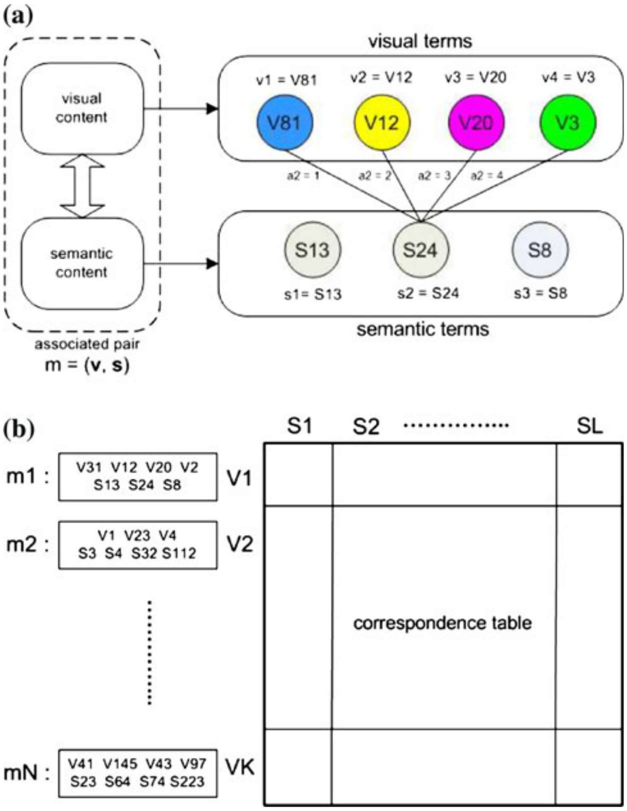


Fig. 3 A framework for solving the missing link problem. **a** Each associated pair m relates the visual terms and semantic terms only roughly. **b** The correspondences are learned from a multimedia collection by using the concurrent occurrences of visual and semantic terms in different associated pairs $(m^1 \ldots m^N)$ and can be provided through a correspondence table

cessed to split it into subparts and to extract descriptive features from each of those subparts. For example, in an image, subparts can be obtained by segmenting it into regions; and in a video, the subparts can be obtained by segmenting it into shots. Similarly, semantic content may be processed to obtain only nouns which are likely to describe visual content, e.g., a named entity detector can be used to extract the names corresponding to faces.

The visual content in an associated pair $m^n = (\mathbf{v}^n, \mathbf{s}^n)$, is then represented as $\mathbf{v}^n = \{v_1^n, \ldots, v_k^n\}$, where each v_i^n is referred to as a *visual term* and k is the number of visual terms in $\mathbf{v}^n$; and the semantic content is represented as $\mathbf{s}^n = \{s_1^n, \ldots, s_l^n\}$, where each s_j^n is referred to as a *semantic term* and l is the number of semantic terms in $\mathbf{s}^n$.

The visual terms v_i^n are from the set $\mathbf{V} = \{V_1, \ldots, V_K\}$, which we refer to as *visual vocabulary*, and similarly the semantic terms s_j^n are from the set $\mathbf{S} = \{S_1, \ldots, S_L\}$, which we refer to as *semantic vocabulary*. We require that both sets consist of discrete elements in order to provide the formulation for a wide range of problems.

Each associated pair provides only loose links between visual terms and semantic terms. Given a single associated pair m^n, a semantic term s_j^n can be associated with any of the visual terms $v_i^n, i = 1 \ldots k$. The correct correspondences between a specific visual term and a specific semantic term can be learned if they co-occur in different associated pairs with different terms. These learned correspondences can then be provided through a *correspondence table* to map visual terms to semantic terms in a new multimedia data.

4.2 Multimedia translation to find correspondences

With this formulation, the problem is very similar to the problem faced in machine translation literature. In machine translation, the words in one language are linked to the words in another language through a lexicon. The typical solution to learn the lexicon is to use an aligned bitext where the rough correspondences at the paragraph or sentence level are known. Brown et al. proposed that having unknown one-to-one correspondences between words, learning the joint distributions of words in two languages from rough correspondences can be formulated as a missing data problem [48]. This relevance is first discussed in [45] and used to link image regions with keywords in annotated image collections. Here, we use this approach to address a more general problem to find the correspondences between any type of visual and semantic data.

Having a set of different associated pairs, as an analogy to the aligned bitext in machine translation, the problem can be reformulated in order to maximize the conditional probability $Pr(\mathbf{s} \mid \mathbf{v})$, which is called the likelihood of translation $(\mathbf{s}, \mathbf{v})$, where $\mathbf{v} = \{v_1 \ldots v_k\}$ is a set of visual terms and $\mathbf{s} = \{s_1 \ldots s_l\}$ is a set of semantic terms. Note that, in this study, we use the direct translation model by discarding the language model.

In the following, we summarize the formulation of Brown et al. adapted to the linking of semantic terms with visual terms, which is referred to as *multimedia translation*. For the details of the formulation and its adaptation to multimedia translation the reader should refer to [45,48].

Among the set of models proposed by Brown et al., we utilized Model 1, which assumes that all connections are equally likely, since there is no order relation among the visual or semantic terms (a detailed discussion of why a more powerful translation model has not been chosen can be found in [52]).

The statistical machine translation approach uses the idea of alignments to indicate connections of words in the source and target strings. Adapting the same idea, for an associated pair with k visual terms and l semantic terms, we denote $a_j = i$ as an alignment to specify that a semantic term s_j in position j in $\mathbf{s}$ is connected to a visual term v_i in position i in $\mathbf{v}$, and $a_j = 0$ if it is not connected to any visual term. Therefore, $\mathbf{a}$ is a vector of integers $a_1 \ldots a_l$, where each a_j can take values from 0 to k.

Considering the alignments, the likelihood of $(\mathbf{s}|\mathbf{v})$ can be written in terms of conditional probability $Pr(\mathbf{s}, \mathbf{a}|\mathbf{v})$, where $\mathbf{a}$ is a random alignment in all possible sets of alignments, as

$$Pr(\mathbf{s}|\mathbf{v}) = \sum_{\mathbf{a}} Pr(\mathbf{s}, \mathbf{a}|\mathbf{v}) \tag{1}$$

Assuming a uniform alignment probability (each alignment being equally probable), for a given set of visual terms, the joint likelihood of a set of semantic terms and an alignment is written as

$$Pr(\mathbf{s}, \mathbf{a}|\mathbf{v}) = \frac{\epsilon}{(k+1)^l} \prod_{j=1}^{l} t(s_j|v_{a_j}) \tag{2}$$

where $t(s_j|v_{a_j})$ is the translation probability of the semantic term s_j given the visual term v_{a_j}, ϵ is a fixed small number used to represent that the number of semantic terms is independent of the set of visual terms in that associated pair, and $(k+1)^l$ is the number of all possible alignments. Then, $Pr(\mathbf{s}|\mathbf{v})$ can be written as in the following by performing sums over all possible alignments:

$$Pr(\mathbf{s}|\mathbf{v}) = \frac{\epsilon}{(k+1)^l} \sum_{a_1=0}^{k} \cdots \sum_{a_l=0}^{k} \prod_{j=1}^{l} t(s_j|v_{a_j}) \tag{3}$$

The goal is to maximize $Pr(\mathbf{s}|\mathbf{v})$ subject to the constraint that for each v

$$\sum_{s} t(s|v) = 1 \tag{4}$$

where v is a particular visual term, and s is a particular semantic term.

Through the introduction of the Lagrange multipliers λ_e and the Kronecker delta function δ, which is equal to one when both of its arguments are the same and equal to zero otherwise, the maximum is achieved for

$$t(s|v) = \lambda_e^{-1} \frac{\epsilon}{(k+1)^l} \times \sum_{a_1=0}^{k} \cdots \sum_{a_l=0}^{k} \sum_{j=1}^{l} \delta(s, s_j)\delta(v, v_{a_j}) \prod_{p=1}^{l} t(s_p|v_{a_p}) \tag{5}$$

where $\sum_{j=1}^{l} \delta(s, s_j)\delta(v, v_{a_j})$ is the number of times v connects to s in $\mathbf{a}$.

Since the translation probabilities appear on both sides, the solution of this maximization problem requires an iterative process and can be achieved using the Expectation Maximization (EM) algorithm as proposed by Brown et al. [48] and as first adapted to the multimedia translation problem in [45]. Here, we summarize this iterative process.

For a training set consisting of a set of aligned pairs $\{(\mathbf{v}^1, \mathbf{s}^1), \ldots (\mathbf{v}^n, \mathbf{s}^n) \ldots (\mathbf{v}^N, \mathbf{s}^N)$, the problem can be turned into finding $t(s|v)$ as

$$t(s|v) = \lambda_e^{-1} \sum_{n=1}^{N} c\left(s|v; \mathbf{v}^n, \mathbf{s}^n\right) \tag{6}$$

where $c(s|v; \mathbf{v}^n, \mathbf{s}^n)$ is the expected number of times that v connects to s in the translation $(\mathbf{s}^n, \mathbf{v}^n)$. Here λ_e^{-1} is used only for normalization.

In Model 1, $c(s|v; \mathbf{v}^n, \mathbf{s}^n)$ can be calculated for each associated pair iteratively as

$$c(s|v; \mathbf{v}^n, \mathbf{s}^n) = \frac{t(s|v)}{\sum_{p=1}^{k} t(s|v_p)} \sum_{j=1}^{l} \delta(s, s_j) \sum_{i=1}^{k} \delta(v, v_i) \tag{7}$$

where $\sum_{j=1}^{l} \delta(s, s_j)$ is the count of s in $\mathbf{s}$ and $\sum_{i=1}^{k} \delta(v, v_i)$ is the count of v in $\mathbf{v}$.

Given an initial estimate for the translation probabilities, the new estimate for $t(s|v)$ is then found using the above equation with the new alignments.

Translation probabilities are stored in a correspondence table. The process starts with some initial values for translation probabilities and assigns the alignments for each associated pair. Assuming that the alignments are given, reconstructing the correspondence table is just a matter of counting. Then, revised translation probabilities are obtained by normalization based on these alignments and the alignments are updated iteratively based on the new values for translation probabilities. Model 1 is proved to have a unique local maximum so that parameters derived for it in a series of EM iterations will converge independent of the initial values.

The computed correspondence probabilities can be used in the following ways:

(i) For predicting the semantic term corresponding to a visual term: in this case, the semantic term with the highest probability given the visual term is chosen.

(ii) For predicting the set of semantic terms given a set of visual terms in a visual content: in this case, the posterior probabilities for all visual terms are marginalized, and a few semantic terms with the highest probabilities are chosen.

(iii) For ranked-retrieval of visual content given a semantic term: in this case, the posterior probabilities for each video content are found as above by marginalizing, but then these posterior probabilities are used to find the relation to a given semantic content.

4.3 Application of multimedia translation on videos

In the following, inspired by the challenges in video data sets, we first re-visit the problem of annotating video keyframes with words through the multimedia translation approach and then introduce new problems to show the applicability of the formulation to various visual and semantic information beyond annotated images.

4.3.1 Linking regions to keywords in annotated keyframes

As shown in Fig. 4, when an *associated pair* is a manually annotated keyframe of a shot, *visual content* refers to the keyframe and *semantic content* refers to the annotation keywords. Keyframes are first segmented into regions, and features are extracted from regions. Then, these features are vector quantized to obtain a vocabulary of *visual terms*. On the other side, a vocabulary of *semantic terms* is built after a pruning step which selects the annotation keywords with frequencies higher than some specified threshold.

When the correspondences are learned, they can be used to automatically predict annotation keywords for unseen keyframes (*automatic annotation*), or to retrieve keyframes based on their relevance to the query words (*ranked retrieval*). Most importantly, since the direct correspondences are

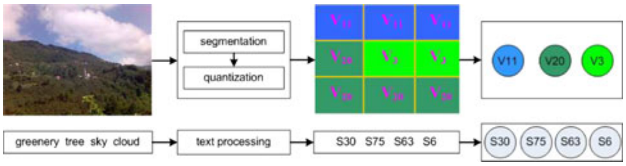
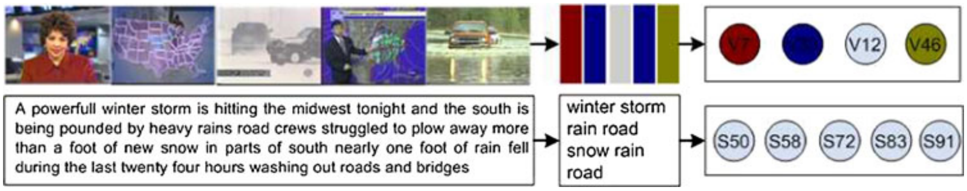


Fig. 4 Linking regions to keywords in annotated keyframes. Regions are mapped to visual terms through segmentation and vector quantization, and annotation keywords are mapped to semantic terms through text-processing techniques. Then the problem is turned into learning the links between visual and semantic terms. While given a single instance learning the links is not possible, the concurrent occurrences in large number of available annotated keyframes provide that information



A powerfull winter storm is hitting the midwest tonight and the south is being pounded by heavy rains road crews struggled to plow away more than a foot of new snow in parts of south nearly one foot of rain fell during the last twenty four hours washing out roads and bridges

winter storm rain road snow rain road

Fig. 5 Linking keyframes to speech transcripts in a video story. The *colored bars* in the middle part represent the quantized visual features of keyframes in the video story which are mapped to visual terms. Similarly, the words in the middle are the words remaining after applying text-processing techniques on the speech transcript text which are mapped to semantic terms. While given a single video story learning the links between visual appearances of keyframes and speech transcript words is not possible, the concurrent occurrences in large number of available video stories provide that information

learned, they can be used to predict labels for the regions (*region labeling*) as an alternative to large-scale object recognition. They can also be used in a setting similar to the query by semantic example method as proposed by [67].

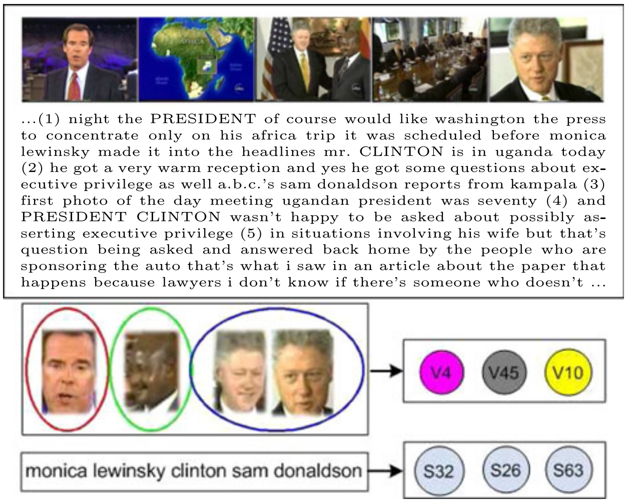
In this setting, the formulation is the same as the approach in [45]. We only re-visit this problem to make the reader familiar with the notation and to prepare a basis for the comparisons with other related approaches (see Sect. 5 for a detailed experimental evaluation). Our main contribution lies in the application of multimedia translation to the following problems.

4.3.2 Linking keyframes to speech transcripts in video stories

In a video story, the speech transcript text is roughly related to the visual information, but the direct correspondences are unknown. In this section, we describe our formulation for this problem and show that learning the direct correspondences in a video story can be achieved by the application of multimedia translation.

As shown in Fig. 5, here the *visual content* in the form of video sequence and the *semantic content* in the form of speech transcript text constitute an *associated pair* within a story. In this study, we define the subparts of the visual content to be the keyframes. Therefore, *visual terms* correspond to the quantized visual representations of the keyframes extracted from the video stories. Since the visual objects and scenes are usually described by nouns, the speech transcript text is processed, and tagging, stemming, and stopword elimination steps are applied to obtain a set of nouns constituting the *semantic terms*.

Making an analogy to the image annotation, in video association problem, regions as being the subparts of images are replaced with keyframes as being the subparts of video stories. Therefore, the counterpart of region labeling process is to predict words for keyframes (*keyframe annotation*), and the counterpart of annotating images is to annotate video stories (*story annotation*).



…(1) night the PRESIDENT of course would like washington the press to concentrate only on his africa trip it was scheduled before monica lewinsky made it into the headlines mr. CLINTON is in uganda today (2) he got a very warm reception and yes he got some questions about executive privilege as well a.b.c.'s sam donaldson reports from kampala (3) first photo of the day meeting ugandan president was seventy (4) and PRESIDENT CLINTON wasn't happy to be asked about possibly asserting executive privilege (5) in situations involving his wife but that's question being asked and answered back home by the people who are sponsoring the auto that's what i saw in an article about the paper that happens because lawyers i don't know if there's someone who doesn't …

monica lewinsky clinton sam donaldson

Fig. 6 Linking faces to names in a video story. First the faces are found using a face detection method. Then similar faces are grouped and mapped to visual terms. Similarly, the names in the text are found and mapped to semantic terms

4.3.3 Linking faces to names in video stories

The formulation for naming faces as finding the correspondences between faces and names is very similar to the one described for the video association problem. It is again a video story that is the *associated pair*. However, in this problem the subparts of visual content are the faces detected in keyframes, and subparts of semantic content are the names appearing in the speech transcript text. The vocabulary of *visual terms* consists of the representatives of face clusters, and the vocabulary of *semantic terms* consists of the names (see Fig. 6).

5 Experimental results

In this section, we present a detailed evaluation of the proposed framework. The experiments are carried out on the TREC Video Retrieval Evaluation (TRECVid) data set provided by NIST [54,55]. Namely, on the TRECVid 2004 corpus which consists of over 150 h of CNN and ABC broadcast news videos, and on the TRECVid 2006 corpus, which consists of 158 h of English, Arabic, and Chinese news videos. The shot boundaries, and the keyframes extracted from each shot are provided by NIST.

The data set is evaluated in the form of two subsets due to the unavailability of some information for the entire data set. In all cases we used 2/3 of the data for training and 1/3 for testing.

The first subset consists of keyframes from the development sets with the available manual annotations. Although the annotations are completed by a collaborative effort of all participants, still this subset is a small portion of the entire data set. This subset is used for the experiments in linking regions of keyframes with annotation words (Sect. 5.2). These experiments are performed only to generate a basis for comparison with other related studies (see Sect. 6).

The second subset is a more challenging one where speech transcript is used as the semantic information. This subset consists of the videos where the segmented video stories are provided and the speech transcript text is available. It is used in the experiments for linking the keyframes of a video story with the speech transcript text (Sect. 5.3), and for linking faces with names (Sect. 5.5).

In this study, we compute the probabilities using the Giza++ tool [68,69], which is part of the Statistical Machine Translation toolkit developed in 1999 at the CLSP at Johns Hopkins University. Specifically Model 1 is utilized and ten iterations are performed in training.

In the following, first we describe the evaluation criteria used, and then present the results for the three problems.

5.1 Evaluation criteria

We first compute standard *recall*, *precision* and F_1 measures for performance evaluation, where

$$\text{recall} = \frac{\text{number of correct predictions}}{\text{number of occurrences in the set}}$$

$$\text{precision} = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$$

$$F_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Then, the *average word recall* and *average word precision* values are obtained by first computing the *recall* and *precision* values for each word (semantic term), and then taking the average over all the words with nonzero recall. These measures are widely used in performance evaluation of existing automatic annotation systems [24,37,38,46].

We use *mean average precision (mAP)* to evaluate the ranked retrieval performance; this is the standard way of evaluating ranked retrieval performance. The *average precision* of a query q, $(AP(q))$, is computed as the sum of the precisions of retrieved relevant images at rank i divided by the total number of relevant images ($relevant(q)$) for the query q. The *mean average precision (mAP)* is then defined as the mean of average precisions (AP) of all queries, N_q.

$$AP(q) = \frac{\sum_{i \in \text{relevant}} \text{precision}(i)}{\text{relevant}(q)}$$

$$mAP = \frac{\sum_{q=1}^{N_q} AP(q)}{N_q}$$

In addition to these standard measures, we use another intuitive measure for annotation performance: *average annotation performance (aap)*. Illustrating with an example for image annotation, if *aap* is specified as 30%; this means 30% of the annotation words predicted for an image in the test set are correct.

$$aap = \frac{\sum_{i=1}^{N} \frac{\text{\# correct words predicted for test image } i}{\text{\# words in test image } i}}{N}$$

5.2 Experiments on linking regions to words in annotated images

In the TRECVid 2004 data set, 114 videos from the development set are manually annotated with a collaborative effort of the participants [70] with 614 keywords, most of which have very low frequency, spelling and format errors. With a preprocessing step, we corrected the errors and excluded the high (higher than 5000) or low (lower than 20) frequencies from the vocabulary, resulting in 115 remaining keywords. There are a total of 28156 keyframes. Similarly, in the TRECVid 2006 data set, LSCOM annotations [71] with 449 concepts are used; 170 concepts remained after preprocessing. There are a total of 31226 keyframes.

On the side of visual content, keyframes are first segmented into regions and then each region is represented by a set of features. In this study, we preferred to use fixed-size grids, and partitioned the keyframes into 5 × 7 grids. We experimented with mean and standard deviation of HSV and RGB values, Canny edge orientation histograms, Gabor filter outputs, and various combinations of these features. We also experimented with the use of salient points for obtaining the visual terms, but the results were worse than the features used, and will not be discussed.

The grids are then transformed into visual terms by vector quantization with k-means using several K values. The choice of K affects the performance of the system; as K increases, there will be more homogeneous terms which have a higher chance of matching a single object or scene, but large K values result in large translation tables, which is costly and noisy. We empirically evaluated several K values ranging from 500 to 5000 with several random partitioning of the

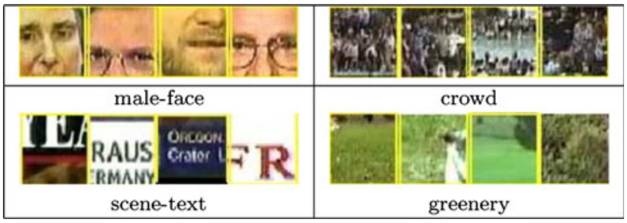


Fig. 7 Examples of matches between visual and semantic terms on TRECVid 2004 data set

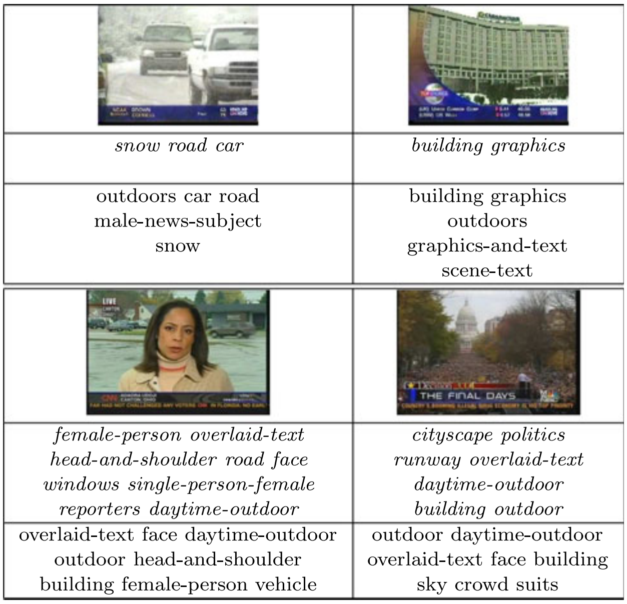


Fig. 8 Image-based auto-annotation examples on (*top*) TRECVid 2004 and (*bottom*) TRECVid 2006 data sets. The manual annotations are shown in *italic*, and the top five predicted words are shown as plain text

data. The best result is obtained by the concatenation of HSV and Gabor features and for $K = 2000$.

The correspondences between visual terms and semantic terms are learned using a training set of 80 videos with 19810 keyframes in TRECVid 2004, and 103 videos with 23950 keyframes in TRECVid 2006. The learned correspondences are used for labeling regions, annotating images, and for ranked retrieval.

Figure 7 shows some visual terms corresponding to some semantic terms with high prediction accuracies. These results show that the links between the semantic and visual terms are correctly learned in most cases. Similarly, Fig. 8 shows some good auto-annotation examples on TRECVid 2004 and 2006 data sets.

Since the quantitative evaluation of region labeling requires the manual labeling of regions, which is costly and subjective, we, instead, use annotation performance as an approximation, since correct correspondences suggests correct annotations.

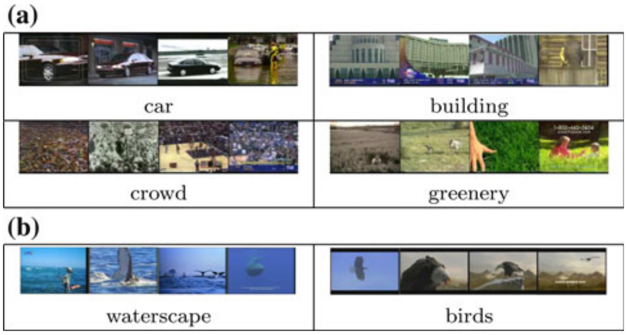


Fig. 9 Image-based ranked query results for some words on **a** TRECVid 2004 and **b** TRECVid 2006 data sets

We compute the annotation performance measures by comparing the predicted annotations with the provided manual annotations. The performance of the system is evaluated using 33 test set videos with 8346 keyframes in TRECVid 2004, and 34 videos with 7276 keyframes in TRECVid 2006. In TRECVid 2004, we obtain an *average annotation performance (aap)* of 28%. Since manual annotations are incomplete (for example while a word corresponding to an object in the scene is correctly predicted, it is not included in the manual annotation) or sometimes even erroneous, the calculated performances are lower than the actual ones. The *average word recall* and *average word precision* values for words with nonzero recall are 0.19 and 0.45, respectively. Out of 115 words used in training, 40 words have nonzero recall in the test set. In TRECVid 2006, the *average annotation performance* is 45%, the *average word recall* and *average word precision* values for the words with nonzero recall are 0.24 and 0.53, respectively. Out of 170 words used in training, 81 words have nonzero recall in the test set. The performance is better in TRECVid 2006, since the annotations are of much higher quality compared to those of TRECVid 2004.

Figure 9 shows query results for some of the highest-ranked words for the TRECVid 2004 and 2006 data sets. For evaluating ranked retrieval results, on TRECVid 2006, queries are performed for all 170 concepts and the relevance of retrieved keyframes is evaluated by comparing them with the LSCOM annotations. When all retrievals are evaluated (for each semantic term, the number of retrieved keyframes is equal to its frequency in the test set), the mAP is 0.28; and when the first 20 retrieved keyframes are evaluated for all terms the mAP is 0.34 for all 170 concepts, and 0.55 for the best 100 concepts. The results show that when annotations are not available, the proposed system can be used effectively for ranked retrieval.

5.3 Experiments on linking keyframes to speech transcripts in news stories

In the experiments using speech transcript text, keyframes of 221 videos from TRECVid 2004 with temporally aligned automatic speech recognition (ASR) transcripts provided by LIMSI [72] were used. ASR is in free text form and requires preprocessing. Therefore, we applied tagging, stemming, and stop-word elimination steps and have used only those nouns that appeared more than 100 but fewer than 1888 times (these rough values are determined by looking at the frequency plot of all the words) to generate a final vocabulary of 297 semantic terms.

The keyframes of the stories are represented by global $18 \times 3 \times 3$ HSV, $4 \times 4 \times 4$ RGB, 17 bin Canny edge orientation, and 1000 bin SIFT keypoint histograms [73,74]. The keyframes are transformed into visual terms by quantizing the features with k-means using several K values ranging from 500 to 5000. The best results are obtained for $K = 3000$. The stories are represented with these visual terms. The story boundaries are provided by NIST. A total of 5768 stories containing 65758 keyframes were used. The number of words corresponding to the stories ranges from 4 to 105, and the average number of words per story is 15.

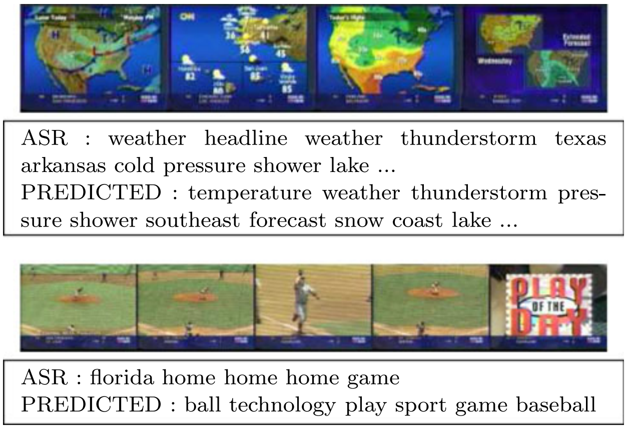
The correspondences between keyframes and speech transcripts are learned using a training set of 155 videos with 4070 stories and 45922 keyframes and used for the annotation of stories and keyframes of shots, and for story-based ranked retrieval.

The system performance is measured using 66 test set videos with 1698 stories and 19836 keyframes. Translation probabilities are used to predict words for the individual shots (Fig. 10) and for the stories (Fig. 11). The results show that the system can predict the correct words, especially for those stories related to `weather`, `sports`, and `economy` that frequently appear in the broadcast news, and which can be represented with the global color and edge features used. The shot auto-annotation examples in Fig. 10 show that even for those shots with no speech transcripts (as in the first two shots), or which have irrelevant words due to the misalignment problem (as in the third shot), the proposed method is able to predict the correct words by learning the links between keyframes and speech transcripts within the story.

The *average annotation performance (aap)* of our system is 23% per story. The *average word recall* and *precision* values for the words with nonzero recall are 0.25 and 0.35, respectively. Out of 297 words used in training, 220 have nonzero recall in the test set. For those 100 words with the highest recall, the *average word recall* and *precision* are increased to 0.40 and 0.45 respectively. Figure 12a shows word recall and precision values for some words.

-	-	*mother doctor*	*game*
flight air bombing airline plane	military america war weapon	sport course baseball	basketball game college play

Fig. 10 Top words predicted for some shots using the ASR outputs on TRECVid 2004. ASR texts aligned with those shots temporally are shown at *top* in italic. Note that, the shots with empty or wrong aligned ASR text can be correctly retrieved using the predicted words



ASR : weather headline weather thunderstorm texas arkansas cold pressure shower lake ...
PREDICTED : temperature weather thunderstorm pressure shower southeast forecast snow coast lake ...

ASR : florida home home home game
PREDICTED : ball technology play sport game baseball

Fig. 11 Example ASR outputs and top words predicted for some TRECVid 2004 stories

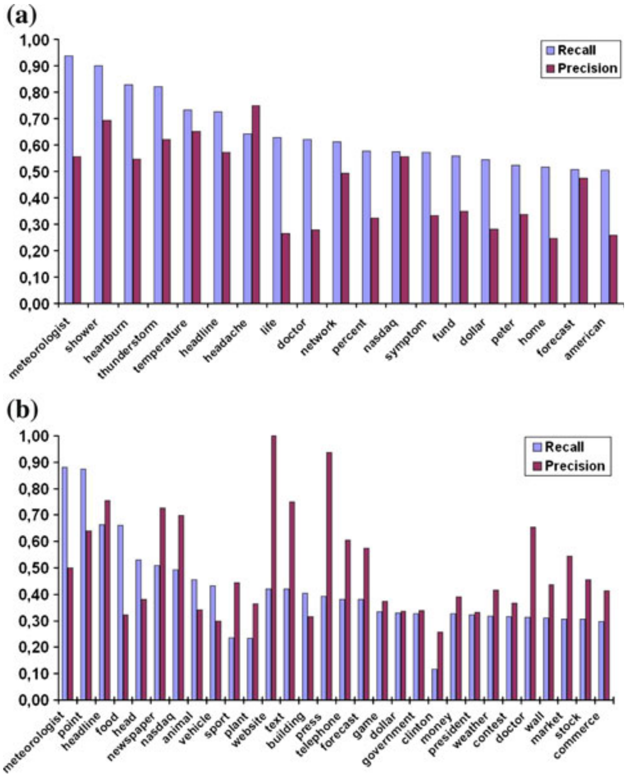


Fig. 12 Word recall and precision values for some words from (**a**) ASR and (**b**) with WordNet hierarchies in ASR on TRECVid 2004

Story-based query results in Fig. 13 show that the proposed system is able to detect the associations between words and scenes/objects. In these examples, the shots within each story are ranked according to the marginalized word posterior probabilities, and the shots matching the query word with

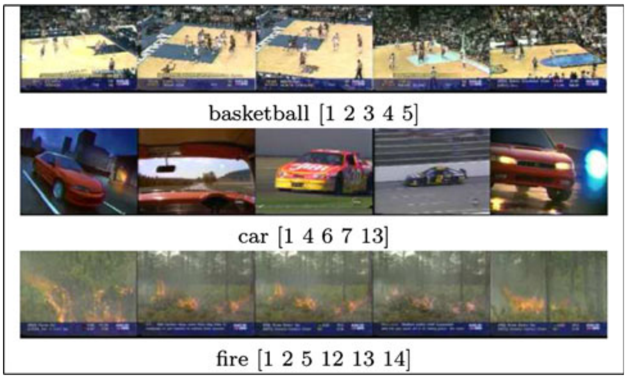


Fig. 13 Story-based ranked query results for ASR on TRECVid 2004. Numbers in *square brackets* show the rank of retrieval

the highest probability are retrieved; a final ranking is done among all shots retrieved from all stories, and all videos and final-ranked query results are returned to the user.

5.4 Use of wordnet

In speech transcript text, a large number of words appear very rarely. When the number of semantic terms is large, associations with visual terms cannot be accurately captured, since the probabilities would be distributed over a large number of terms. On the other hand, most of the words share similar semantic meanings. If these semantic relationships were captured, then the associations could be learned more effectively. Ontologies are helpful in creating semantic relationships, but their creation is costly and moreover they may be incomplete. In this study, we use WordNet [75] to automatically create an ontology; we take the words in the WordNet hierarchy instead of original ASR words as the semantic terms.

We have incorporated the WordNet hierarchies into both manual annotations and ASR text. First, for each word, those words that are in the upper level WordNet hierarchy are added to the vocabulary (e.g., `sport` is added for `football`, `baseball`, `basketball`, `tennis`, etc.). Second, part meronyms (part-whole relationships) are added (e.g., `leaf` is added for `plant`). Finally, the vocabulary is constructed by excluding words with frequencies which are too high or too low.

Incorporating WordNet hierarchies enhances performance in several ways. First, vocabulary size is increased with the addition of new words, leading to a better representation; the increase was 162 words for manual annotations, and 566 for ASR. Second, there was a considerable increase in the *average annotation performance*, *average word recall* and *average word precision* values, and the number of words with nonzero recall. For manual annotations, the number of words with nonzero recall increased to 70, and *average word recall* and *average word precision* increased to 0.25 and 0.50, respectively. For ASR, the number of words with nonzero recall increased to 435, and average annotation performance increased to 35%. The addition of new ontological words, as well as improvements in word recall and precision values are reflected in the word recall-precision graph (Fig. 12b).

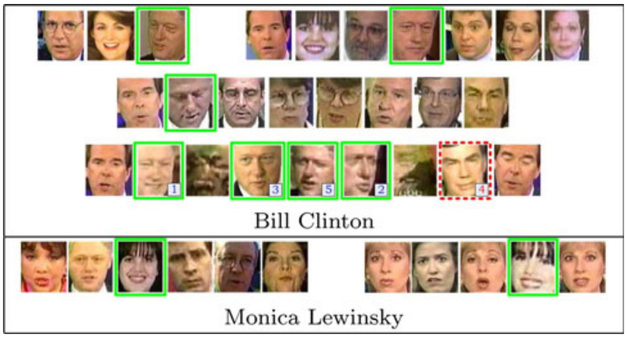


Fig. 14 Aligning faces with correct names within stories. Each group contains the faces detected in a story in the order they appear. Boxes around faces indicate the retrieved faces which are associated with the names (*green* correct, *dashed red* wrong), and numbers at the bottom right show the rank of retrieval

5.5 Experiments on linking faces to names

For face naming, the same set of videos for story-based alignment is used (TRECVid 2004). Faces in the keyframes are detected, and represented with 6 × 5 grids of color (mean-std of HSV, RGB) and texture (9 bin Canny edge orientation histogram, Gabor filter outputs, 500 bin SIFT keypoint histogram) features. The faces are transformed into visual terms by quantizing the features with k-means using several K values (50, 100, 200). HSV concatenated with Gabor using $K = 200$ gives the best results. 92 person names, only 25 of which have a frequency greater than 100, are extracted from the speech transcripts manually, since our focus is not on named entity detection. The correspondences between faces and names are learned using a training set of 155 videos with 2070 stories and 6428 faces and then used for aligning faces with names within the stories.

Figure 14 presents some samples in which the faces in each story are ranked according to the probability of their association with the queried name and retrieved successfully. Irrelevant shots would be retrieved in a solely text-based system since the query names are not aligned with the correct faces. For instance, in the second row of Fig. 14, the name *Clinton* appears in a previous shot which does not even contain a face.

6 Comparative experiments

In this section, we present the results of comparative experiments with several approaches. First, we evaluate four approaches proposed for auto-annotation from the second and the third categories. Then, we compare multimedia translation method with the traditional classifier-based methods by focusing on multi-class Support Vector Machines (SVM) on TRECVid 2006 high-level feature extraction task.

6.1 Comparison of auto-annotation approaches

As stated previously, methods other than those of the third category suffer from not directly learning one-to-one correspondences and may thus not be suitable for solving problems in other domains, such as for the recognition of objects and faces. On the other hand, the methods in all categories are applicable to the image annotation problem and can therefore be compared on this basis.

To understand the advantages and disadvantages of the multimedia translation method (Translation) over other methods, we selected one method from the second group, namely the Cross Media Relevance Model (**CMRM**) proposed by Jeon et al. [36] and two other methods from the third category, namely the *Cooccurrence* method proposed by Mori et al. [44], and the **SvdCos** method proposed by Pan et al. [46] and compared them based on their image annotation performances. All four models work on the discrete representation of the visual data.

Since the implementations of these methods are not publicly available, we implemented them ourselves according to the descriptions in their papers. We used the same data set used in Sect. 5.2, with 115 semantic terms and 2000 visual terms obtained by the concatenation of HSV and Gabor features extracted from 5×7 grids and quantized with k-means.

The results are summarized in Table 1. In $case_1$, the number of words predicted per image is the same as the number of actual annotation words, while in $case_2$, 5 words are predicted per image. The SvdCos, Translation, and Co-occurrence methods can predict a smaller number of words compared to the CMRM method, but with higher precision. Fig. 15 shows the F_1 scores for words with best prediction performance for $case_2$. Considering the F_1 scores for words with nonzero recall, the Co-occurrence method has the highest performance, while CMRM has the lowest. According to the *average annotation performance*, all four methods are equally successful, with the Translation method performing slightly better.

In $case_1$, the number of words with nonzero recall is largest for CMRM (81), and smallest in Cooccurrence (18).

When we compare only among these 18 words, the following results are obtained, as shown in Table 2: CMRM has the highest *average word recall* (0.45), Cooccurrence has the highest *average word precision* (0.49) but lowest *average word recall* (0.27), and Translation has the highest F_1 score (0.39) by a small margin.

Table 1 On automatic annotation task, performance comparison of four methods: SvdCos, CMRM, Cooccurrence, and Ours (translation)

Measure	SvdCos	CMRM	Cooc	Ours
$case_1$				
avg word rec	0.23	0.22	0.27	0.19
avg word prec	0.45	0.21	0.49	0.45
F_1	0.30	0.22	0.35	0.27
# words	42	81	18	40
aap	0.27	0.26	0.26	0.28
$case_2$				
avg word rec	0.29	0.28	0.35	0.29
avg word prec	0.26	0.11	0.35	0.22
F_1	0.28	0.16	0.35	0.25
# words	50	89	25	44
aap	0.48	0.43	0.45	0.49

aap average annotation performance, *rec* recall, and *prec* precision. The number of words indicates, out of 115 words, how many words have nonzero recall. Data set: TRECVid 2004

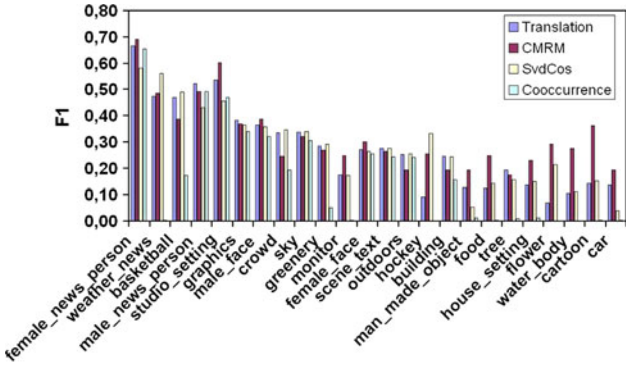


Fig. 15 Comparison of our method (translation) with Coocurrence, SvdCos, and CMRM using F_1 values at five predicted words for some highest-ranked words on TRECVid 2004 data set

Table 2 Performance comparison for the best 18 words, for the setup in Table 1

Measure	SvdCos	CMRM	Cooc	Ours
avg word rec	0.38	0.45	0.27	0.39
avg word prec	0.31	0.33	0.49	0.39
F_1	0.34	0.38	0.35	0.39

Depending on the type of application, one may choose to predict more words with low accuracy or fewer words with higher accuracy. For instance, a web search engine may prefer to predict a smaller number of frequently searched concepts with higher accuracies. However, there is a trade-off between the number of words predicted and word-prediction performance. Predicting too few words would be undesirable, as in the case of the Co-occurrence method. Therefore, the Translation and SvdCos methods are more balanced compared to the other two.

Another issue in the selection of one of the annotation methods is the use of context information. As stated earlier, CMRM learns the joint distribution of words and images within a context. This is an advantage of CMRM in learning rare words and thus predicting more words, since the context does not change much in the TRECVid data set for the training and test examples. However, it becomes a big disadvantage when the context changes since the words can no longer be predicted.

To gain more insight into how context affects annotation performance, we have created a small, noise-free artificial data set. We have constructed a training set by creating permutations of ten concepts. We assume that each image contains two concepts with visual terms V_i and V_j, and with two corresponding semantic terms S_i and S_j. That is, the training set includes the following matching tuples $(\{V_1, V_2\}, \{S_1, S_2\})$, $(\{V_1, V_3\}, \{S_1, S_3\})$, ... $(\{V_9, V_{10}\}, \{S_9, S_{10}\})$. We further assume that there is a rare concept which only appears with concept 1 but nothing else, adding another tuple $(\{V_1, V_{11}\}, \{S_1, S_{11}\})$ to the training set. In the test set, there are five other unseen visual terms co-occurring with the visual terms in the training set. Now we consider two cases in the test set: $(\{V_1, V_{15}\})$ and $(\{V_{11}, V_{15}\})$, where V_1 and V_{11} are the visual terms already observed in the training set while V_{15} is not.

We do not expect any system to predict a semantic term S_{15} corresponding to visual term V_{15}, but the predictions of semantic terms corresponding to V_1 and V_{11} tell much about the context dependency. With our experiments on CMRM and Translation methods, we observe that the Translation method can predict the correct semantic terms in both cases, whereas the CMRM method can predict for V_1, but not for V_{11}, meaning that the rare words cannot be predicted by the CMRM method if they are not appearing in the same context of the training samples. On the other hand, the Translation method can predict any word without depending on the context, since it learns direct correspondences between visual and semantic terms.

As a real-case example, the foreign minister of each country is most likely to be seen with his prime minister in the news. Therefore, a context-dependent approach is likely to couple the face of a prime minister to the face of his foreign minister, since it would increase its performance. Now, consider the performance of such a system on a news photo taken from a summit of foreign ministers from all over the world. In this case, the context-dependent approach would suffer, since the context is completely different from the one it was trained in, as nicely demonstrated by the toy problem considered above.

The problem can be overcome with the use of larger volumes of data with the assumption that in the ideal case, every pair appearing in the data, however, should be considered carefully.

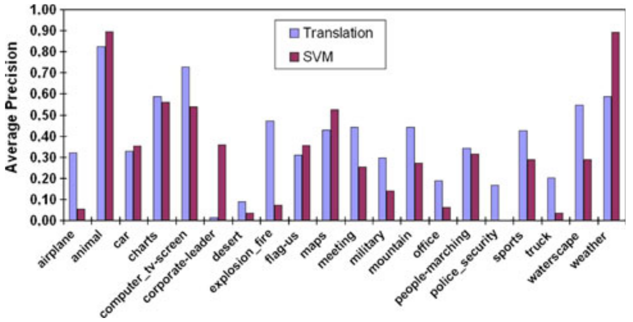


Fig. 16 Comparing translation with SVM-based method on TRECVid 2006 high level feature extraction task

6.2 Comparison with classifier-based approaches

To compare the performance of translation approach with the classifier-based methods, we trained our system using the same set of high level concepts used in TRECVid 2006 high-level feature extraction task [55]. In TRECVid 2006, 39 concepts (LSCOM-lite annotations) are used in high-level feature extraction task and 20 of these are evaluated. Most of the submissions for high-level feature extraction task use SVM to learn the concepts. Participants used the whole development set for training. The performance of each system was evaluated by NIST on TRECVid 2006 test set by visually inspecting the submissions. Since we used the development set for both training and testing due to the availability of manual annotations, our training and test sets are not the same as the ones used in TRECVid evaluation. We trained a multi-class SVM as described in [76] on the same data as we used in Translation by using the multi-class SVM tool in [77]. We performed single-word queries for all 39 concepts using Translation and multi-class SVM methods and retrieved as many keyframes as each concept appears in the test set, and computed the average precision for each concept. Figure 16 shows average precision values for 20 concepts that were evaluated in TRECVid 2006. Translation approach performs considerably better in most of the concepts. High performance figures for some concepts (e.g., for animal, weather, computer-tv-screen) are most probably due to duplicates or near duplicates existing in training and test sets.

7 Summary and discussion

In this study, we attack the semantic gap problem in large annotated data sets where multimedia data appear with semantic information and formulate it as the missing correspondences between the low-level multimedia data and the semantic content. After reviewing the studies in image annotation which integrate visual data to semantics using loosely labeled data sets and discussing their relevancy to the semantic gap problem, we proposed that the semantic gap can be addressed as the translation of multimedia data to semantics.

Our main contribution is threefold: (i) expanding the domain of association problems from images to video sequences and presenting new problems, (ii) commonly formulating these problems within the same framework, and (iii) adaptation of translation method to a wide range of problems going beyond image annotation.

The multimedia translation method has originally been developed within the machine learning literature and adapted to learning in multimedia problems in [45,39]. The novelty of this study does not lie in the translation method itself, but in the adaptation of the method to new problems. In particular, we have described problems in which this proposed framework and the translation method can be applied: linking regions of keyframes with annotation keywords, linking keyframes in a video story to speech transcripts, and linking faces to names.

With extensive experimental studies on the TRECVid 2004 and TRECVid 2006 corpora which consist of over 300 h of news videos, we show that novel applications including the automatic annotation of video keyframes, video story annotation, large-scale object and face recognition, and as a result of all, better retrieval on large-scale video datasets are possible.

On TRECVid 2006, we obtained better results in annotation performance compared to TRECVid 2004 most probably due to better manual annotations, but worse results in story annotation experiments due to the low performance of the automatic story boundary detection.

We also conducted additional experiments and compared the performance of the proposed method with other auto-annotation and classifier-based systems and showed its advantages over other methods.

In this study, we used the statistical machine translation approach to find the associations. However, we argue that any other method, with the requirement that it learns direct correspondences between visual terms and semantic terms, could also be used in place of the translation method.

We should also note that, although we focus on visual information, there is no limit in the choice of multimedia content, which could be anything such as audio, range data, sensor outputs, symbols, etc. Similarly, in this study, semantics is represented in the form of text, but it can also be in some other form.

Acknowledgments This research is partially supported by TÜBİTAK Career grant number 104E065 and grant number 104E077, and by the European COST292 Action on Semantic Multimodal Analysis of Digital Media. We are grateful to Ann-Marie Thornton and Kirsten Ward for proofreading.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
2. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **2**(1), 1–19 (2006)
3. Hare, J., Lewis, P., Enser, P., Sandom, C.: Mind the gap: another look at the problem of the semantic gap in image retrieval. In: Multimedia Content Analysis, Management and Retrieval, SPIE, vol. 6073, San Jose, California, USA (2006)
4. Chang, S., Hsu, A.: Image information systems: where do we go from here? IEEE Trans. Knowl. Data Eng. **4**(5), 431–442 (1992)
5. Rui, Y., Huang, T., Chang, S.: Image retrieval: current techniques, promising directions, and open issues. J. Vis. Commun. Image Represent. **10**(4), 39–62 (1999)
6. Snoek, C., Worring, M.: Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools Appl. **25**(1), 5–35 (2005)
7. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
8. Zhao, R., Grosky, W.I.: Narrowing the semantic gap: improved text-based web document retrieval using visual features. IEEE Trans. Multimedia **4**(2), 189–200 (2002)
9. Benitez, A., Chang, S.-F.: Semantic knowledge construction from annotated image collections. In: IEEE International Conference on Multimedia and Expo, vol. 2, pp. 205–208 (2002)
10. Chang, S.-F., Manmatha, R., Chua, T.-S.: Combining Text and audio-visual features in video indexing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, Philadelphia, PA, March, pp. 1005–1008 (2005)
11. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
12. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 878–885 (2005)
13. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **2**, II-264–II-271 (2003)
14. Wu, L., Hu, Y., Li, M., Yu, N., Hua, X.-S.: Scale-invariant visual language modeling for object categorization. IEEE Trans. Multimedia **11**(2), 286–294 (2003)
15. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (2006)
16. Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T.: A thousand words in a scene. IEEE Trans. Pattern Anal. Mach. Intell. **29**(9), 1575–1589 (2007)
17. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: IEEE CVPR 2004, Workshop on Generative-Model Based Vision (2004)
18. Caltech 101 Dataset Homepage [Online]. Available: http://www.vision.caltech.edu/Image_Datasets/Caltech101
19. Griffin, G., Holub, A., Perona, P.: Caltech-256 Object Category Dataset, California Institute of Technology, Tech. Rep. 7694 (2007)
20. Everingham, M., Zisserman, A., Williams, C., Gool, L. V., Allan, M., Bishop, C., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J.: The 2005 PASCAL Visual Object Classes Challenge. In: Selected Proceedings of the First PASCAL Challenges Workshop, LNAI, Springer-Verlag (2006)
21. The PASCAL Visual Object Classes Homepage [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC

22. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**(1–3), 157–173 (2008)
23. LabelMe Homepage [Online]. Available: http://labelme.csail.mit.edu
24. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image retrieval. Pattern Anal. Mach. Intell. **29**(10), 1802–1817 (2007)
25. Getty Images [Online]. Available: http://www.gettyimages.com
26. Flickr Photo Sharing Service [Online]. Available: http://www.espgame.org/gwap
27. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: ACM Conference on Human Factors in Computing Systems (CHI 2004), pp. 319–326 (2004)
28. The ESP Game [Online]. Available: http://www.flickr.com
29. Yahoo! News [Online]. Available: http://news.yahoo.com
30. Kender, J. R., Naphade, M. R.: Visual concepts for news story tracking: analyzing and exploiting the NIST TRECVID video annotation experiment. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1174–1181 (2005)
31. Enser, P., Sandom, C.J., Hare, J., Lewis, P.: Facing the reality of semantic image retrieval. J. Document. **63**(4), 465–481 (2007)
32. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: The 15th International Conference on Machine Learning, pp. 341–349 (1998)
33. Argillander, J., Iyengar, G., Nock, H.: Semantic annotation of multimedia using maximum entropy models. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Philadelphia, PA, USA, March 18–23, pp. 153–156 (2005)
34. Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, San Diego, June, pp. 163–168 (2005)
35. Li, J., Wang, J.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1075–1088 (2003)
36. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: 26th Annual Int. ACM SIGIR Conference, Toronto, Canada, July 28–August 1, pp. 119–126 (2003)
37. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: 17th Annual Conference on Neural Information Processing Systems, vol. 16, pp. 553–560 (2003)
38. Feng, S., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. In: International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1002–1009 (2004)
39. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D.A., Blei, D., Jordan, M.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)
40. Blei D., Jordan, M.I.: Modeling annotated data. In: 26th Annual International ACM SIGIR Conference, Toronto, Canada, July 28–August 1, pp. 127–134 (2003)
41. Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. In: International Conference on Computer Vision, vol. 2, pp. 408–415 (2001)
42. Hofmann, T., Puzicha, J.: Statistical Models for Co-occurrence Data, AI Memo 1625, CBCL Memo 159, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, MIT, Tech. Rep., February (1998)
43. Monay, F., Gatica-Perez, D.: PLSA-based image auto-annotation: constraining the latent space. In: ACM International Conference on Multimedia, October, pp. 348–351 (2004)
44. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management (1999)
45. Duygulu, P., Barnard, K., Freitas, N., Forsyth, D.A.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: 7th European Conference on Computer Vision, vol. 4, Copenhagen Denmark, May 27–June 2, pp. 97–112 (2002)
46. Pan, J.-Y., Yang, H.-J., Duygulu, P., Faloutsos, C.: Automatic image captioning. In: The 2004 IEEE International Conference on Multimedia and Expo, vol. 3, Taipei, Taiwan, June, pp. 1987–1990 (2004)
47. Carbonetto, P., de Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. In: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11–14, pp. 350–362 (2004)
48. Brown, P., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
49. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
50. Barnard, K., Duygulu, P., Guru, R., Gabbur, P., Forsyth, D.: The effects of segmentation and feature choice in a translation model of object recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Madison, Wisconsin, June, pp. 675–682 (2003)
51. Jeon, J., Manmatha, R.: Using maximum entropy for automatic image annotation. In: 3rd International Conference on Image and Video Retrieval, Ireland, July 21–23, pp. 24–32 (2004)
52. Virga P., Duygulu, P.: Systematic evaluation of machine translation methods for image and video annotation. In: The 4th International Conference on Image and Video Retrieval, Singapore, July 20–22, pp. 174–183 (2005)
53. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image indexing. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1802–1817 (2007)
54. TREC Video Retrieval Evaluation [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid
55. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid, In: 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330 (2006)
56. Smeaton, A., Over, P., Kraaij, W.: High level feature detection from video in TRECVid: a 5-year retrospective of achievements. In: Divakaran, A. (ed.) Multimedia Content Analysis, Theory and Applications. Springer, Berlin (2008)
57. Ghoshal, A., Ircing, P., Khudanpur, S.: Hidden Markov models for automatic annotation and content based retrieval of images and video. In: The 28th International ACM SIGIR Conference, Salvador, Brazil, August 15–19, pp. 544–551 (2005)
58. Duygulu, P., Hauptmann, A.: What's news what's not? Associating News videos with words. In: The 3rd International Conference on Image and Video Retrieval (CIVR 2004), Ireland, July 21–23, pp. 21–23 (2004)
59. Wactlar, H., Hauptmann, A., Witbrock, M.: Informedia News-On Demand: Using Speech Recognition to Create a Digital Video Library, CMU Technical Report, CMU-CS-98-109, Tech. Rep. (1998)
60. Gross, R., Baker, S., Matthews, I., Kanade, T.: Face recognition across pose and illumination. In: Li, S.Z., Jain, A.K. (eds.) Handbook of Face Recognition. Springer, Berlin, pp. 193–216 (2004)
61. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. **35**(4), 399–458 (2003)
62. Yang, J., Chen, M.-Y., Hauptmann, A.: Finding Person X: correlating names with visual appearances. In: International Conference on Image and Video Retrieval, Ireland, pp. 270–278 (2004)
63. Berg, T., Berg, A.C., Edwards, J., Maire, M., White, R., Teh, Y.-W., Learned-Miller, E., Forsyth, D.: Faces and names in the news.

In: IEEE Conference on Computer Vision and Pattern Recognition (2004)

64. Ozkan, D., Duygulu, P.: A graph based approach for naming faces in news photos. In: IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1477–1482 (2006)
65. Satoh, S., Kanade, T.: Name-It: Association of face and name in video. In: IEEE Conference on Computer Vision and Pattern Recognition (1997)
66. Baştan, M., Duygulu, P.: Recognizing objects and scenes in news videos. In: The International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, **40071**, pp. 380–390 (2006)
67. Rasiwasia, N., Vasconcelos, N.: Bridging the semantic gap: query by semantic example. IEEE Trans. Multimedia **9**(5), 923–938 (2007)
68. Giza++: Training of statistical translation models [Online]. Available: http://www.fjoch.com/GIZA++.html
69. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **1**(29), 19–51 (2003)
70. Lin, C. -Y., Tseng, B. L., Smith, J. R.: Video Collaborative annotation forum: establishing ground-truth labels on large multimedia datasets. In: NIST TREC-2003 Video Retrieval Evaluation Conference, Gaithersburg, MD, November (2003)
71. Naphade, M., Curtis, J., Hauptmann, A., Kennedy, L., Hsu, W., Chang, S.-F., Smith, J.: Large-scale concept ontology for multimedia. IEEE Multimedia **13**(3), 86–91 (2006)
72. Gauvain, J., Lamel, L., Adda, G.: The LIMSI broadcast news transcription system. Speech Commun. **37**(1-2), 89–108 (2002)
73. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
74. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV) (2003)
75. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: an online lexical database. Int. J. Lexicogr. **3**, 235–244 (1990)
76. Tang, J., Lewis, P.: A study of quality issues for image auto-annotation with the corel data-set. IEEE Trans. Circuits Syst. Video Technol. **17**(3), 384–389 (2007)
77. Joachims, T.: Multi-class support vector machine [Online]. Available: http://svmlight.joachims.org/svm-multiclass.html