Signal Processing: *Image Communication* 28 (2013) 311–322

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Selection of local features for visual search

Gianluca Francini*, Skjalg Lepsøy, Massimo Balestri

Telecom Italia, Via Reiss Romoli, 274, IT-10148 Torino, Italy

ARTICLE INFO

Available online 19 November 2012

Keywords:
Content-based image retrieval
Image search
Visual search
Compact descriptors
MPEG

ABSTRACT

In a compact descriptor for visual search only a limited number of local features may be included. The estimated probability for correct match between keypoints provides a good criterion for selection of a subset.

1. Introduction

Pairwise matching and retrieval are two fundamental tasks for a system of visual search. The former regards automated verification of whether two images depict the same objects or scene. The latter regards the search and discovery of images contained within a large collection that depict the same objects or scene as those depicted by a query image. Both tasks may be initiated by a user equipped with a mobile terminal that can take pictures and send and receive information over a network.

In order to minimize the amount of data sent over the network and reduce latency time, the terminal should be able to extract from the picture and transmit only the information that is essential to the matching or retrieval task at hand [1]. This extraction is the goal of programs that compute *compact descriptors for visual search*.

A descriptor is a sequence of bits that represent information about an image, so a compact descriptor is one that contains few bits. We shall be concerned with compact descriptors that are constrained to be shorter than certain lengths, for example 512 bytes, 1 KB, 2 KB or similar.[1] Such descriptors may contain information about the image as a whole (provided by global features) and information about the image in certain parts (provided by local features.) As the length is constrained, it is important to pack only the information that is most relevant to the given task into the descriptor.

Here we shall consider only local features, such as SIFT [3], SURF [4], and CHoG [5]. A local feature is a vector of values whose elements characterize a neighborhood around a certain point (keypoint or interest point) in an image, in correspondence to an image detail. The local features allow matching of keypoints in one image to keypoints in another image; two keypoints are matched because their local features are similar.

Matched points in two images often correspond to the same points in the depicted scene, but there is no guarantee that they do. By estimating how likely any feature is to be correctly matched, we may eliminate the least likely ones and pack only the most promising features into the compact descriptor. This process will be called *feature selection*. The probability of correct match will be called the *relevance*.

We shall select such features based on certain quantities that are available as output of the feature *detection* process. In the case of the SIFT feature, these quantities are: location of the keypoint, the scale, the absolute value of the detected extremum, and the orientation [3]. Each of these quantities conditions the probability that the feature may be matched correctly. We use a training set to

* Corresponding author.
E-mail address: Gianluca.Francini@telecomitalia.it (G. Francini).

[1] These lengths are considered in the standardization activity 'Compact Descriptors for Visual Search' (CDVS) under MPEG-7 [2].

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2012.11.002

estimate these conditional probabilities. For each quantity, we also propose an explanation of why it should have an influence on the relevance of a feature. Finally, a relevance score is obtained by multiplying the conditional probabilities.

The next section introduces the probabilistic grounds for the method of feature selection by relevance. The following section provides the details of the training procedure. Then, the method is applied to determine how the relevance of SIFT features depends on each of the quantities output by the detection process. Finally, we report on the gain brought by the feature selection to pairwise matching of images belonging to different categories.

2. Feature relevance

The feature selection assigns a positive value to any feature, as a function of some quantities related to the keypoint detection. We will call these *characteristic* quantities.

The basic assumption is that the features of correctly matched keypoints are different from the features of wrongly matched keypoints (as far as the statistics of the characteristic quantities are concerned). As will be seen, this difference in behaviour is confirmed by experiments, and it is furthermore found to be quite consistent across various datasets, so that parameters obtained by analyzing a training set are applicable with success to a test set.

We let the set of characteristic quantities for the mth feature in an image be denoted by s_m. The function value will be denoted by $r(s_m)$ (for *relevance*). The relevancies for the features in one image will be sorted such that

$$r(s_{m_1}) \geq r(s_{m_2}) \geq \cdots \geq r(s_{m_M}).$$

Only the L most relevant features $m_1,\ldots,m_L$ are kept, such that the descriptor length remains below the given constraint.

The relevance function will be embodied by an estimator for the probability that a feature may be matched correctly to some feature in some unknown image, provided that the characteristic quantities are quantized to some region. Suppose that we use only one characteristic quantity γ in order to quantify the probability that the local feature will be matched correctly, and suppose that we have observed that γ lies within some region G. The conditional probability for correct matching (denoted by $c=1$ as opposed to $c=0$) is then

$$P(c=1\,|\,\gamma \in G) = \frac{P(c=1 \cap \gamma \in G)}{P(\gamma \in G)}. \tag{1}$$

As stated above, we will use more than one quantity in order to estimate the probability of correct match. We make the assumption that these quantities $\alpha,\ldots,\omega$ are conditionally independent given c (akin to the 'naïve Bayesian classifier' assumption [6]), leaving the equations quite simple. Under this assumption the probability of correct match is proportional to a product

$$\begin{aligned} &P(c=1\,|\,\alpha \in A,\ldots,\omega \in \Omega)\\ &\quad \propto P(c=1\,|\,\alpha \in A)\ldots P(c=1\,|\,\omega \in \Omega)\\ &\quad = \frac{P(c=1 \cap \alpha \in A)}{P(\alpha \in A)}\cdots\frac{P(c=1 \cap \omega \in \Omega)}{P(\omega \in \Omega)}. \end{aligned} \tag{2}$$

In Eq. (2), the regions $A,\ldots,\Omega$ represent cells of quantizers for each quantity $\alpha,\ldots,\omega$. If we let the quantizer for quantity α have the cells $A_1,\ldots,A_{K_\alpha}$, and so forth until the quantizer for ω which will have the cells $\Omega_1,\ldots,\Omega_{K_\omega}$, then the quantizers and Eq. (2) together define a *relevance function r*

$$\begin{aligned} &\alpha \in A_{k_\alpha},\ldots,\omega \in \Omega_{k_\omega}\\ &\quad \Downarrow\\ &\quad r(\alpha,\ldots,\omega) = \frac{P(c=1 \cap \alpha \in A_{k_\alpha})}{P(\alpha \in A_{k_\alpha})}\cdots\frac{P(c=1 \cap \omega \in \Omega_{k_\omega})}{P(\omega \in \Omega_{k_\omega})}. \end{aligned} \tag{3}$$

3. Training

The training estimates the conditional probabilities of correct match for all quantizer cells and all quantities

$$\begin{aligned} &\frac{P(c=1 \cap \alpha \in A_1)}{P(\alpha \in A_1)},\ldots,\frac{P(c=1 \cap \alpha \in A_{K_\alpha})}{P(\alpha \in A_{K_\alpha})},\\ &\quad\vdots\\ &\quad\frac{P(c=1 \cap \omega \in \Omega_1)}{P(\omega \in \Omega_1)},\ldots,\frac{P(c=1 \cap \omega \in \Omega_{K_\omega})}{P(\alpha \in \Omega_{K_\omega})}. \end{aligned}$$

The training proceeds by automated supervised learning and is carried out on matching pairs of images (both images in a pair depict the same object).

Training set: For producing the training set, each image pair undergoes the following operations:

1. Detection of keypoints and extraction of local features in both images. For each feature a set of characteristic quantities is computed $s=\{\alpha,\ldots,\omega\}$.
2. Matching of local features within each image pair. All features that were detected but not matched are labeled as $c=0$. The other features are processed in step 3.
3. Automated labeling of each match as correct ($c=1$) or incorrect ($c=0$). This may be done by a geometric consistency check with RANSAC [7,8] or DISTRAT [9].

We let B, C denote a pair of images. The steps 1 and 2 produce a list of M matches such that $s_m(B)$, $s_m(C)$ represent the characteristic quantities for the matched pair number m. Step 3 labels this pair with label c_m. Each non-matched feature (with characteristic quantities $\overline{s}$) is labeled with $c=0$ in step 2. The part of the training set