

# 영상 클래스별 중요 특징 가중에 의한 영상 검색 방법

유동근, 박채훈, 최유경, 권인소  
한국과학기술원 전기 및 전자 공학과  
e-mail : {saviola02, like24w, ykchoi82, iskweon}@kaist.ac.kr

## Image Retrieval by Important Feature Weighting for Each Class

Donggeun Yoo, Chaehoon Park, Yukyung Choi, In So Kweon  
Dept. of Electrical Engineering, KAIST

### 요약

이 논문에서는 영상 검색(image retrieval) 및 영상 부류(image categorization)을 위하여 영상을 기술할 때 영상의 클래스(class)별로 서로 다른 주요 특징량(feature)에 가중치를 주는 방법론을 제안한다. 기존에 연구되어온 영상의 특징량 벡터에 가중치를 주는 방식은 모든 영상 클래스에 대하여 동일하게 가중치를 적용하기 때문에 영상이 클래스별로 서로 다른 특징량이 중요하다는 성질을 이용할 수 없다. 영상이 클래스별로 서로 다른 특징량이 중요하다는 성질을 이용하기 위하여 영상의 클래스별로 특징량 벡터에 서로 다른 가중치 벡터(weight vector)를 학습하였다. 그 후 질의 영상(query image)이 입력되면, 기존의 영상 검색 프레임워크(framework)를 통해 데이터베이스(database)로부터 미리 정의된 서브 클래스(sub-class)의 수에 해당하는 영상 부 집합(subset)을 만들었다. 그리고 영상 부 집합의 특징량 벡터들에 클래스별로 각각 학습된 가중치 벡터를 적용하여 특징량 벡터들 간의 거리를 다시 계산하여 리랭킹(re-ranking)하였다. 이 방법론을 UKBench Dataset에 적용하여 실험을 해보았으며 가중치를 주기 전과 비교하였을 때 더 높은 정확도를 보였다.

### 1. 서론

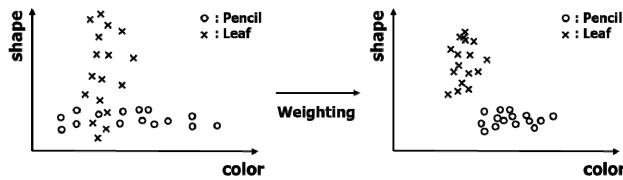
영상 검색은 컴퓨터 비전의 중요한 분야 중 하나이다. 그동안의 정보 검색은 검색하고자 하는 내용을 문자로 입력하여 원하는 결과를 얻어내는 문자 검색(text retrieval)을 중심으로 많은 발전을 이루었다. 오늘 날 멀티미디어 기술의 급속한 발달로 폭발적인 양의 영상 데이터를 제공받을 수 있게 되면서 그 자료를 효율적으로 검색하는 기술이 중요해졌다. 하지만 검색하고자 하는 내용이 문자로 표현할 수 없는 영상인 경우, 영상을 입력으로 받아 그 내용을 분석하고 의미 있는 검색 결과를 얻어내는 영상 검색의 경우 아직 정확한 결과를 보장하지 못하고 있어 컴퓨터 비전 분야에서 활발히 연구되고 있다.

영상 검색을 위해서는 영상의 수치적인 기술이 필요한데, 이때 영상 자체의 특징 정보인 색상과 모양, 질감[1], 관심 영역(interesting region)[2, 3] 등 다양한 방식으로 영상을 기술하게 된다. 여러 방법론들 중 가장 보편적으로 쓰이는 것은 지역 특징에 기반한 시각 어휘(visual vocabulary)[4]를 생성하여 문자 검색 분야의 기본적인 프레임워크를 영상 검색에 적용하는 것이다. 시각 어휘를 생성하는 방법은 영상의 영역을 기술하는 것부터 시작한다. 어파인(affine), 크기(scale), 회전(rotation), 병진(translational)과 같은 기하 변환(geometric transformation)에 대하여 불변적이고, 조명, 노이즈, 흐려짐(blur)과 같은 측광학적 변환(photometric transformation)에 대하여 강인한 특성을



(그림 1) 캔, 연필 클래스의 중요한 시각단어

보이는 영역 기술 방법[2, 3]을 통해 영역을 기술한다. 그리고 학습 영상들로부터 기술한 기술자들을 모아 양자화를 통해 시각 어휘를 생성하게 된다. 하나의 영상은 양자화를 통해 생성된 각 시각 단어가 얼마나 출현하는지를 히스토그램(histogram)으로 표현하여 영상 전역을 기술하게 된다. 하지만 검색하고자 하는 영상 데이터베이스의 크기가 커질수록 영상 클래스간의 구별성이 저하되기 때문에 이를 해결하기 위한 방법 또한 연구되고 있다. 영상 클래스간의 구별성을 높이기 위해 양질의 시각 단어를 생성하거나 시각 단어의 수를 크게하기도 하고[5], 구별성이 큰 특징량을 선별적으로 이용하기도 하며[6, 7], 같은 위치의 부 영역(sub-region)에서 매칭(matching)된 시각 단어에 가중치를 부여하기도 한다.[7, 8]



(그림 2) 두 클래스의 가중치 적용 전 특징량 공간과 가중치 적용 후 특징량 공간

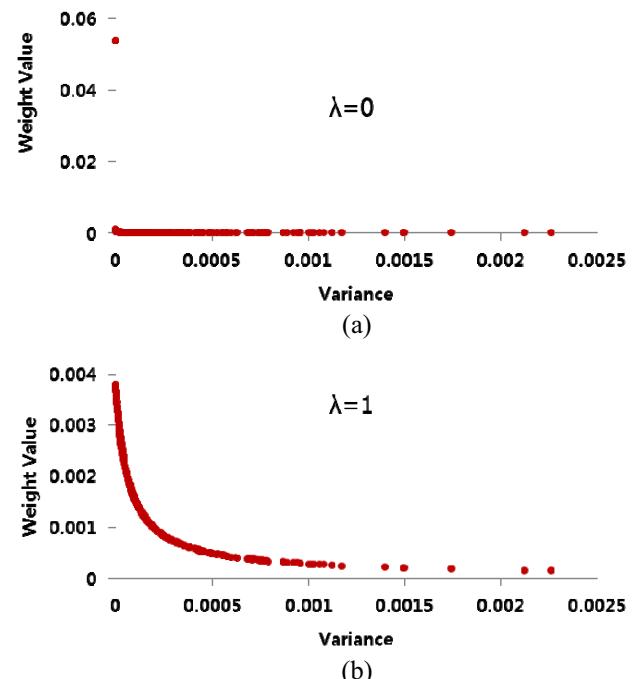
영상 클래스간 구별성을 높이기 위하여 연구되어 온 이와 같은 방법들은 모든 클래스의 영상들에 대하여 같은 기준을 적용할 뿐 서로 다른 클래스의 영상은 중요한 특징량이 서로 다르다는 것을 고려하지 않는다. 하지만 (그림 1)의 예시와 같이 음료수 캔 영상의 중요한 시각 단어와 연필 영상의 중요한 시각 단어가 다른 것처럼 영상 클래스별로 중요한 특징량이 다르기 때문에 본 논문에서는 이를 고려한 영상 클래스별 특징량 가중치 학습 방법과 이를 적용하여 영상 검색을 하는 방법을 제안한다.

## 2. 영상 클래스별 가중치 학습

한 클래스의 학습 영상들에 대하여 작은 분산을 갖는 특징량은 상대적으로 중요한 특징량이고, 큰 분산을 갖는 특징량은 상대적으로 덜 중요한 특징량이라는 기본 가정으로부터 영상 클래스별로 가중치 벡터(weight vector)를 구하는 알고리즘을 제안한다.

(그림 2)의 가중치를 주기 전 특징량 공간(feature space)에서와 같이 펜 클래스와 나뭇잎 클래스는 각 특징량에 대하여 서로 다른 분포를 보일 것이다. 펜 클래스에서는 펜의 모양이 대부분 긴 모양을 하기 때문에 모양(shape)을 나타내는 특징량에 대해서는 좁은 분산을 보이고, 펜의 색깔은 매우 다양하기 때문에 색(color)을 나타내는 특징량에 대해서는 넓은 분산을 보일 것이다. 이와 마찬가지로 나뭇잎 클래스에 대해 보면, 나뭇잎은 매우 다양한 모양을 갖기 때문에 모양을 나타내는 특징량에 대해서는 넓은 분산을 가질 것이고, 대부분의 나뭇잎은 녹색이기 때문에 색을 나타내는 특징량에 대해서는 좁은 분산을 가질 것이다. 즉, 모양을 나타내는 특징량에 대하여 좁은 분산을 갖는 펜 클래스는 모양이 중요한 특징량이고, 색을 나타내는 특징량에 대하여 좁은 분산을 갖는 나뭇잎 클래스는 색이 중요한 특징량이다. 이 논리에 따라 펜 클래스에는 색에 작은 가중치를 주고 나뭇잎 클래스에는 모양에 작은 가중치를 준다면 (그림 2)의 가중치를 준 후의 특징량 공간처럼 각 클래스의 영상들이 서로 모이게 될 것이다. 즉, 각 영상 클래스별로 가중치 벡터를 학습하는 방법은 각 영상 클래스의 학습 데이터들이 최대한 각 클래스에 대하여 서로 모이도록 하는 가중치 벡터를 찾는 문제가 된다.

각 영상 클래스의 학습 데이터들이 최대한 서로 모이도록 하는 가중치 벡터를 찾기 위하여 비용 함수(cost function)를 정의하고 그 비용 함수를 최소화하는 가중치 벡터를 구한다.



(그림 3) 불이익 항(penalty term) 추가 전의 분산-가중치 그래프(a)와 추가 후의 분산-가중치 그래프(b)

$$\text{Cost}(\mathbf{w}_c) = \sum_i \sum_j \left\{ \sum_d \left( w_{c,d} (x_{d,i} - x_{d,j}) \right)^2 \right\}^{\frac{1}{2}} \quad (1)$$

$$\text{s.t. } \sum_d w_{c,d} = 1$$

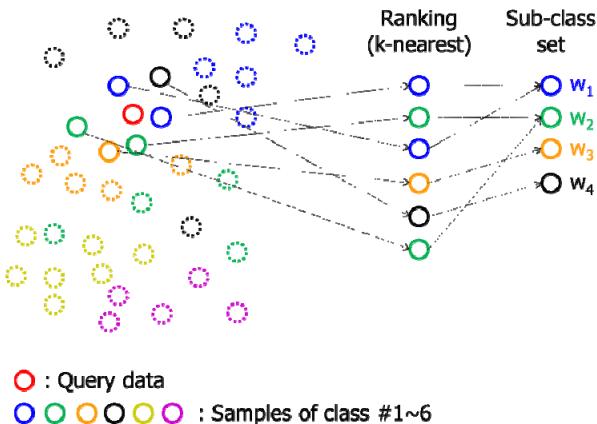
식 (1)은 한 클래스( $c$ )의 서로 다른 모든 영상 쌍( $i, j$ )에 대하여 특징량 벡터( $\mathbf{x}$ )의 각 차원( $d$ )별로 가중치 벡터( $\mathbf{w}_c$ )가 적용된 유clidean 거리(Euclidian distance)의 합을 나타낸다. 이 때 가중치 벡터의 성분들을 합이 1이 되도록 제약을 주었다. 대용량 영상 검색(large scale image retrieval)에서는 데이터들을 역파일(inverted file) 구조를 사용하기 때문에, 모든 벡터들이  $l^2$  정규화( $l^2$ -normalization) 된 상태에서 질의 영상이 입력되면 질의 영상의 벡터와 데이터베이스의 벡터를 내적하여 코사인 유사도(cosine similarity)를 구한다. 즉 코사인 유사도를 사용할 때 더 정확한 클래스별 가중치를 계산하기 위하여  $l^2$  정규화 되어 있는 특징 벡터들 사이의 거리를 호의 길이(arc length)로 정의하였고, 이를 비용 함수에 적용하였다.

$$\text{Cost}(\mathbf{w}_c) = \sum_i \sum_j \cos^{-1} \left\{ \sum_d (w_{c,d} \cdot x_{d,i}) \times (w_{c,d} \cdot x_{d,j}) \right\} \quad (2)$$

$$\text{s.t. } \sum_d w_{c,d} = 1$$

식 (2)는 한 클래스( $c$ )의 서로 다른 모든 영상 쌍( $i, j$ )에 대하여 특징량 벡터( $\mathbf{x}$ )의 각 차원별( $d$ )로 가중치 벡터( $\mathbf{w}_c$ )가 적용된 호의 길이의 합을 나타낸다.

$$\hat{\mathbf{w}}_c = \arg \min_{\mathbf{w}_c} \text{Cost}(\mathbf{w}_c) \quad (3)$$



(그림 4) 질의 영상으로 부터 4개의 서브 클래스 모음(sub-class set)을 만드는 방법

식 (1), (2)와 같은 비용함수를 식 (3)과 같이 각 클래스 별( $c$ ) 가중치 벡터( $w_c$ )를 적용하여 최소화하면 각 영상 클래스의 학습 특징량 벡터들이 서로 가장 가깝게 모일 수 있는 가중치 벡터를 얻게 된다. 하지만 식 (1)과 (2)와 같이 단지 특징 벡터들 사이 거리의 합만을 비용 함수로 설계했을 때에 (그림 3)의 (a) 그래프와 같이 이를 최소화 시키는 가중치 벡터는 가장 작은 분산을 갖는 특징량 차원에만 큰 가중치를 주고 나머지 차원에 대해서는 0에 가까운 가중치를 주는 문제가 발생한다. 특정 한 차원에만 가중치가 크게 할당되면 영상 클래스간의 구별에 하나의 차원만 영향을 미치게 되므로 오히려 영상 클래스별 구별성이 현저히 떨어지게 된다. 이 문제를 해결하기 위해 가중치 벡터의  $l^2$  크기( $l^2$ -norm)를 불이익 항(penalty term)으로써 비용 함수에 추가한다.

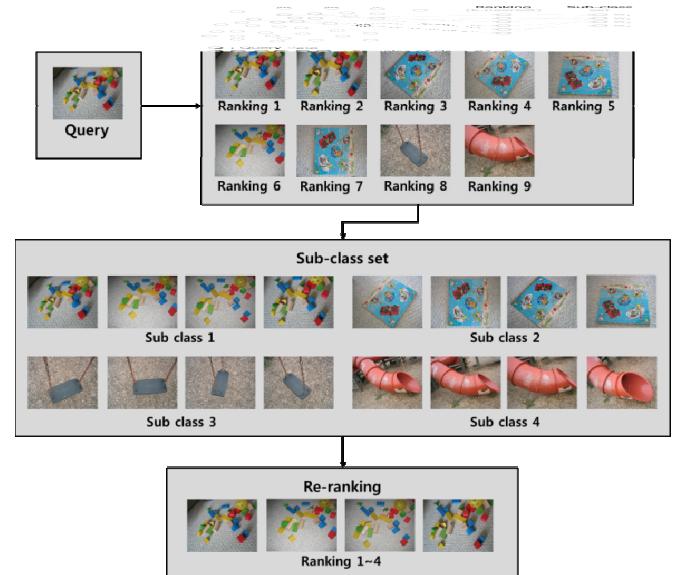
$$Cost(w) = \sum_i \sum_j dist(x_i, x_j, w) + \lambda \cdot \|w\|_2 \quad (4)$$

$$s.t. \sum_d w_d = 1$$

식 (4)와 같이 식 (1), 식 (2)에서 만든 특징량 벡터들 간의 가중치 적용된 거리( $dist(x_1, x_2, w)$ ) 합에 불이익 항에 적절한 가중치( $\lambda$ )를 곱해 주면 (그림 3)의 (b) 그래프와 같이 작은 분산을 갖는 특징량 차원에 대해서는 큰 가중치가, 큰 분산을 갖는 특징량 차원에 대해서는 작은 가중치가 할당된다.

### 3. 가중치를 적용한 검색 방법

영상 클래스별로 학습된 가중치 벡터를 적용하여 영상 검색을 하는데에는 하나의 문제점이 있다. 질의 영상의 클래스를 모르기 때문에 어떤 영상 클래스의 가중치 벡터를 적용할지 모른다. 이 문제를 해결하기 위해 가중치를 적용하지 않은 상태로 데이터베이스로부터 영상 부집합을 만든 후 부집합의 영상에 가중치를 적용해 리랭킹하는 방법을 제안한다.

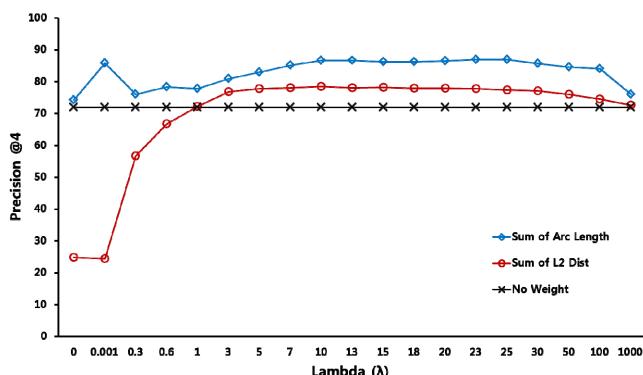


(그림 5) 4개의 서브 클래스 모음(sub-class set)을 만드는 실제 예

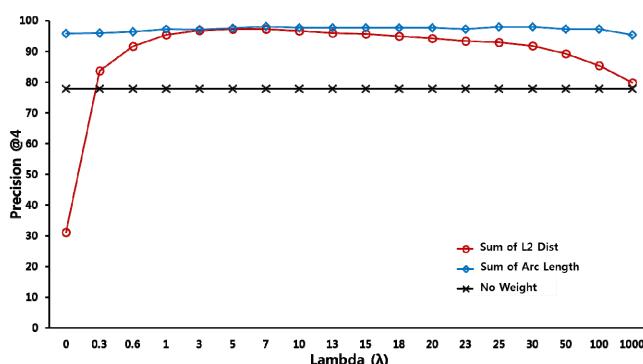
(그림 4)에서와 같이 가중치를 적용하지 않은 상태로 질의 영상으로 부터 미리 정의된 수 만큼의 서로 다른 클래스가 나타날 때 까지 코사인 유사도를 이용하여 가장 가까운 순서대로 영상을 검색한다. 그 후 검색된 영상의 클래스를 서브 클래스 모음(sub-class set)을 만든 후 그 클래스에 해당하는 모든 영상을 가져와 영상 부집합을 만든다. 그 후 질의 영상 특징량 벡터에 각 부집합 영상이 가진 가중치 벡터를 적용하여 다시 질의 영상 특징량과 부집합 영상들 간의 거리를 계산하여 리랭킹한다. (그림 5)는 서브 클래스 모음을 만드는 실제 예를 보여준다. 여기서는 4개의 서로 다른 클래스의 영상이 찾아질 때 까지 가중치를 적용하지 않은 채 영상을 검색한 후 그 클래스에 속하는 모든 영상들을 가져와 각 클래스별 가중치를 적용하여 리랭킹 하였다.

### 4. 실험 결과 및 분석

본 논문에서는 제안한 영상 검색 알고리즘의 성능을 평가하기 위하여 한 영상 클래스 당 4 장의 영상으로 구성된 UKBench dataset[5]의 전체 데이터 중 250 개의 영상 클래스를 가져와 총 1000 장의 영상으로 실험하였다. 영상의 특징점 추출 및 기술은 SIFT[2]를 사용하였으며, 500 개의 시각 어휘를 생성하기 위해  $k$ -평균 군집화( $k$ -means clustering) 알고리즘을 사용하였다. 영상 특징량 벡터는 두 가지 방식을 사용하였다. 첫 번째는 시각 단위 히스토그램(bag-of-words histogram)[4]만을 한 영상을 나타내는 특징량 벡터로 사용하였기 때문에 한 장의 영상은 500 차원의 특징량 벡터로 표현된다. 두 번째 방식은 500 차원의 시각 단위 히스토그램과 영상의 채널별로 가로와 세로를 각각 8 픽셀로 축소한 192 차원의 축소 영상(tiny image)[9], RGB 를 각 채널별로 평균한 3 차원의 벡터를 이어 붙여 한 영상의 특징량 벡터로 사용하였다.



(그림 6) 시각 단어 히스토그램만을 사용한 불이익 항의 비중에 따른 정확도 실험



(그림 7) 시각 단어 히스토그램, 축소 영상, RGB평균 벡터를 사용한 불이익 항의 비중에 따른 정확도 실험

이 세 종류의 특징량 벡터의 영향을 같게 하기 위해 각각  $l^2$  정규화 하여 이어 붙인 후 전체를 다시  $l^2$  정규화 하였다. 각 클래스의 영상 중 한 장의 영상을 질의 영상으로 하여 총 250 장의 질의 영상을 만들었고, 코사인 유사도를 거리로 사용하였으며, 리랭킹 결과 가장 가까운 4 장의 영상 중 참영상 개수의 평균을 백분율[5]로 계산해 평가했다. 또한 가중치를 주지 않은 것과 유클리드 거리의 합을 비용 함수로 사용해 가중치를 준 것, 호의 길이의 합을 비용 함수로 사용해 가중치를 준 것의 정확도를 비교하였다.

(그림 6)는 시각 단어 히스토그램만을, (그림 7)은 세 종류의 특징량 벡터를 사용하여 불이익 항의 비중에 따른 정확도를 나타낸다. 적절한 불이익 항의 비중을 주면 유클리드 거리의 합을 비용 함수로 사용하였을 때, 영상에 가중치를 주지 않은 것 보다 높은 정확도를 보였다. 불이익 항에 아주 작은 비중을 주면 특정 특징량 차원에 거의 모든 가중치가 가해지기 때문에 정확도가 현저히 떨어진다. 하나의 차원에 거의 모든 가중치가 가게 되면 영상간의 구별에 영향을 미치는 차원이 거의 한 차원에 의해 결정나기 때문이다. 호의 길이의 합을 비용 함수로 사용하였을 때에는 코사인 유사도와 같은 거리를 사용한 것이기 때문에 더 높은 정확도를 보였고, 대부분의 불이익 항의 비중에 대해서도 정확도가 더욱 향상되었다.

## 5. 결론

본 논문에서는 영상의 클래스별로 중요한 특징량이 서로 다른다는 점을 영상 검색에 적용하기 위하여 영상 클래스별 가중치 벡터를 학습하는 방법과 학습된 가중치 벡터를 이용해 영상검색을 하는 방법을 제안하였다. 제안된 방법으로 각 클래스에 중요한 특징량에 다른 가중치를 주어 성능을 개선하였다. 하지만 가중치 벡터 학습 알고리즘이 한 클래스의 학습 영상 수가 많아질 경우 오버피팅(over-fitting)되기 쉽고, 시각단어의 수가 커질수록 학습시간이 기하급수적으로 늘어나 대용량 영상검색에 적용하기 힘들다는 문제가 있다. 이를 해결하기 위하여 연구를 진행하고 있고, 앞으로 더 개선해야 할 것이다.

## 6. 감사의 글

This research was supported by MKE (The Ministry of Knowledge Economy), Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA (National IT Industry Promotion Agency).

## 참고문헌

- [1] Varma, M., Zisserman, A. "A statistical approach to texture classification from single images," International Journal of Computer Vision, Vol.62, pp.61-81, 2005.
- [2] Lowe, D. "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, Vol.60, No.2 pp.91-110, 2004.
- [3] Mikolajczyk, K. and Schmid, C. "Scale & affine invariant interest point detectors," International Journal of Computer Vision, Vol.60, No.1, pp.63-86, 2004.
- [4] Josef Sivic and Andrew Zisserman. "Video Google: A Text Retrieval Approach to Object Matching in Videos," In IEEE International Conference on Computer Vision, 2003.
- [5] D. Nister, H. Stewenius, "Scalable recognition with a vocabulary tree," In IEEE Conference on Computer Vision and Pattern Recognition, 2006.
- [6] Y. Freund, R. Iyer, R. Schapire, and Y. Singer. "An efficient boosting algorithm for combining preferences," Journal of Machine Learning Research, Vol.4, pp.933-969, 2003.
- [7] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang and Lei Zhang. "Spatial-Bag-of-Features," In IEEE Conference on Computer Vision and Pattern Recognition, 2010.
- [8] Lazebnik, S., Schmid, C., Ponce, J. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," In IEEE Conference on Computer Vision and Pattern Recognition, 2006.
- [9] A. Torralba, R. Fergus, W. T. Freeman. "80 million tiny images: a large dataset for non-parametric object and scene recognition," Vol.30, No.11, pp.1958-1970, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.