MPEG-7 오디오 특징을 이용한 감성기반 음악검색

임지혜*, 이준환**
*전북대학교 컴퓨터공학과
**전북대학교 컴퓨터공학과
e-mail:jhlim@jbnu.ac.kr

Emotion-Based Music Retrieval using MPEG-7 Audio Descriptors

Jee-Hye Lim*, Joon-Whoan Lee**
*Dept of Computer Engineering, Chonbuk University
**Dept of Computer Engineering, Chonbuk University

요 약

음원의 디지털화와 다양한 디지털 기기의 보급으로 인해 사용자는 더욱 쉽게 많은 양의 음악을 접할 수 있게 되었다. 많은 양의 음원중에서 사용자 개개인의 성향에 맞는 음악을 검색하기 위해 내용기반 음악검색과 감성기반 음악검색 방법 등이 제안되고 개발되고 있다. 본 논문에서는 감성기반 음악검색 방법에서 다차원 벡터 형태의 MPEG-7 저수준 오디오 서술자들의 중요도를 결정하기 위한 새로운 방법을 제안하였다. 제안된 방법은 한 쌍의 대립되는 감성을 대표하는 음악들의 유사성을 다차원 서술자의 관점에서 측정한다. 그리고 이 유사관계를 러프 근사화와 군집 내/군집 간의 유사성 비율을 이용하여 서술자의 중요성을 결정하는데 사용한다. 이 중요성을 바탕으로 결정된 가중치는 여러 개의 오디오 서술자들의 유사성을 총체화하여 감성기반 음악검색에 이용된다.

1. 서론

음원의 디지털화와 다양한 디지털 기기의 보급으로 인해 사용자는 더욱 쉽게 많은 양의 음악을 접할 수 있게 되었다. 초기의 음원 제공 서비스는 단순히 음원을 많이 확보하고 제공하는 서비스에 중점을 두었지만 현재는 많은 양의 음원중에서 사용자 개개인의 성향에 맞는 음악을 제공하는 서비스에도 많은 관심을 기울이고 있다.

하지만 전통적인 음악 정보라 할 수 있는 음악가, 장르, 제목, 앨범 타이틀 등을 이용하여 음원을 검색하는데는 한계가 있다. 따라서 검색 질의를 음악 본연의 내용으로 구성하는 내용기반 검색 시스템이 활발히 연구되고 있다[1][2][3]. 감성기반 음악검색은 내용기반 음악검색에서 더 진보된 형태의 음악검색 방법이다. 사용자가 원하는 감성을 기반으로 음악을 검색해 주는 방법이다.

감성기반 음악검색을 하기 위해서는 우선 감성을 정의하고 표현해야 한다. 본 연구에서는 많은 연구에서 감성을 표현하는데 사용되는 Thayer의 감성 모델을 사용하였다.

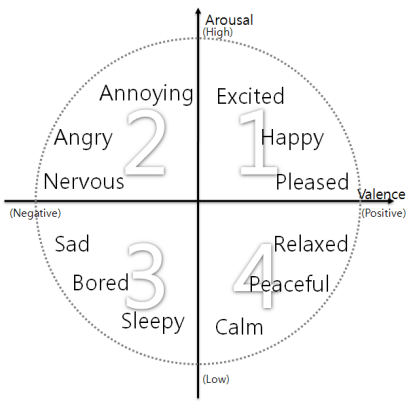
감성은 주관적이므로 특정 감성에 해당하는 음악은 사용자마다 다르다. 따라서 특정 감성에 해당하는 음악을 여러개 제시하고 이들 중 자신의 감성에 가장 잘 맞는 음악을 사용자가 선택한 뒤 그와 유사한 음악을 내용기반 검색을 하여 결과로 제시하도록 하였다.

본 연구에서는 음악의 오디오 속성 값 추출과 검색에 내용기반 멀티미디어 검색 분야에서 표준화된 MPEG-7 저수준 오디오 서술자(Low-Level Audio Descriptor)들을 속성으로 사용한다.

본 연구에서 사용하는 MPEG-7 서술자들은 다차원 벡터이다. 이러한 다차원 벡터들 중에 어떠한 서술자가 어떤 감성의 판정에 얼마나 중요한가를 판정하는 문제 즉 가중치 결정 문제는 대단히 중요하다. 본 연구에서는 퍼지 유사도를 기반으로 러프 근사화와 군집 내/군집 간 유사성 비교를 이용한 두 가지 가중치 결정방법을 제안하였다. 제안된 방법은 신송이 등이 제안한 방법과 동일한 구조에서 감성기반 검색 결과 우수함을 실험적으로 입증하였다.

2. 감성 공간 및 오디오 서술자

가. 음악의 감성공간



(그림 1) Thayer의 감성 모델

본 연구에서 감정을 추출하기 위해 사용하는 Thayer 모델은 감성을 Arousal과 Valence 두 축을 사용하여 표현한다. Arousal은 감성의 강도를 의미하며 이 값이 커질수록 흥분한(Excited) 기분을, 작을수록 고요한(Calm) 기분을 의미한다. Valence는 감성의 긍정적/부정적 정도를 나타내며 값이 클수록 긍정적 감성을 나타내며, 값이 작을수록 부정적 감성을 나타낸다. 본 연구에서는 서로 대립을 이루는 6쌍 총 12개의 감성 형용사를 사용하였다[4]. 정의된 형용사는 (그림 1)에 나타나있다.

나. MPEG-7 오디오 서술자

본 연구에서는 MPEG-7 저수준 오디오 서술자들의 (비)유사성 측도를 활용한다. 저수준 오디오 서술자는 오디오세그먼트(AudioSegment : 오디오를 시간적으로 분할한 일부분)에서 사용되도록 고안된 간단하고 낮은 복잡성을 가진 서술자의 집합으로 구성된다[5]. 저수준 오디오 서술자 중 실험에 사용된 서술자들을 간단히 설명하면 아래와 같다. (괄호안의 숫자는 서술자의 차원이다.)

· AudioFundamentalFrequency(300) : 오디오 신호의 기본 주파수를 서술.

· AudioSignature(288) : 오디오의 강력한 자동식별을 위해 필요한, 유일한 콘텐츠 식별자를 제공하기 위해 고안된 오디오 신호의 축약된 표현.

· AudioSpectrumBasis(580) : 스펙트럼 서술을 저차원 표현으로 사영(Projection)하기 위한 기본 함수를 포함. 스펙트럼이 차원을 줄이는 것은 오디오 세그먼트에 대한 특징의 통계 정보를 축약 적으로 표현하기 때문에 자동 분류 응용에서 중요함.

· AudioSpectrumCentroid(900) : 로그 주파수 전력 스펙트럼의 무게중심(center of gravity)을 서술.

· AudioSpectrumEnvelope(48600) : 로그 주파수 축을 가진 스펙트럼의 시간의 열로써 오디오파형의 단기 전력 스펙트럼을 서술.

· AudioSpectrumFlatness(4800) : 오디오 신호의 단기 전력 스펙트럼의 편평성을 서술. 잡음이나 충격(impulse) 신호에 해당하는 편평한 모양, 음조가 있는 분분에서는 높은 변화.

· HarmonicSpectralSpread(1) : Harmonic Spectral Centroid로부터 얻은 하모닉 피크들의 진폭이 가중치로 부여된 편차.

· SpectralCentroidType(1) : 파워 스펙트럼(Power Spectrum) 상에서 주파수들의 에너지를 가중치로 부여한 평균.

MPEG-7 오디오 서술자들은 다차원 벡터로 표현되기 때문에 유사성을 측정하기 위해서는 벡터 계산이 필요하다. 본 연구에서는 표준안의 권고대로 유클리디안 거리를 이용한 유사도 측정방법을 사용하였다. 유사도와 거리는 역수관계이므로 유사도를 구하는 가장 간단한 방법은 유클리디안 거리의 역수를 취하는 것이다. 하지만 벡터간의 스케일이 고려되지 않는다는 문제점이 있다.

$$similarity = \frac{1}{1+\log(D+1)} \qquad (1)$$

D : 두 벡터간 유클리디안 거리

식 (1)은 log 함수를 이용해 벡터간의 스케일에 따른 거리 값 차이를 보정했다. 본 연구에서는 식 (1)을 이용해 오디오 서술자간의 유사도를 구했다.

3. 퍼지 유사도 기반 가중치 결정방법

MPEG-7에서 제공하는 모든 오디오 속성을 다 사용하면 계산량이 많고 검색 속도가 느려지는 단점이 있다. 따라서 어떤 속성이 중요한지, 어떤 속성이 어떤 감성을 평가하는데 중요한지를 판별한다. 그리고 이를 감성검색에서 유사성측도 설계에 이용하여 단점을 보완할 수 있다.

이들 속성의 중요도는 대립하는 두 감성의 대표곡들을 얼마나 잘 분류하는가에 따라 결정된다. 본 논문에서는 가중치 결정 방법을 제안하고 이를 이용해 감성기반 음악검색을 하였다. 사용된 가중치 결정방법은 퍼지 유사도를 기반으로한 러프 근사화 방법과 군집 내/군집 간의 유사도의 비율을 이용한 방법이다.

가. 퍼지 유사관계를 이용한 러프 근사화(rough approximation)

1982년 폴란드의 Pawlak이 처음 제창한 이래 고전적인 러프집합 이론[6]은 구별 불가능(indiscernibility)관계의 상위(upper) 및 하위(lower) 근사화 방법을 활용하였다. 논의집합 U의 부분집합 X에 대해 부분특징 $B \subset A$의 퍼지 유사성 관계 R_B의 λ레벨의 하한 근사화는 다음과 같이 정의된다.

$$\underline{R_B^{\lambda}}(X) = \{x \in X : R_B^{\lambda}(x) \subseteq X\} \qquad (2)$$

하한 근사화의 개념을 이용하여 $\{X_i : i = 1,2,...,r(d)\}$의 R_B^{λ}-긍정영역(positive region)은 다음과 같이 정의된다.

$$POS(R_B^{\lambda}, \{d\}) = \bigcup_{i=1}^{r(d)} \underline{R_B^{\lambda}}(X_i) \qquad (3)$$

Stepaniuk는 퍼지 유사관계를 활용하여 각 특징의 가중치를 결정하는 방법을 제안하고 있다.

$$SRC(R_A^{\lambda}, \{d\}, a) = \frac{\left|POS(R_A^{\lambda}, \{d\}, a)\right| - \left|POS(R_{A-\{a\}}^{\lambda}, \{d\}, a)\right|}{|U|} \qquad (4)$$

결국 특징 a를 특징집합 A로부터 제거했을 때 긍정영역이 감소하는 정도가 특징 a의 가중치를 결정한다. 실제 가중치는 $SRC(R_A^{\lambda}, \{d\}, a)$를 정규화하여 결정할 수 있다.

$$w_a^{\lambda} = \frac{SRC(R_A^{\lambda}, \{d\}, a)}{\sum_a SRC(R_A^{\lambda}, \{d\}, a)} \qquad (5)$$

나. 군집 내/ 군집 간 (intra-inter cluster)의 유사성의 비율을 이용한 방법

어떤 특징 a관점에서 $\{X_1,...,X_{r(d)}\}$의 결정 클래스들 내부의 객체들은 서로 유사하며 클래스 상호간의 객체들에서는 유사하지 않다면 이는 중요한 특징일 수 있다.

특징 a관점에서 결정클래스 X_i의 내부적인 유사성의 평균치와 클래스 X_i와 X_j 상호간의 유사성의 평균을 각각 다음식과 같이 정의하자.

$$AIR_a(X_i)=\frac{1}{|X_i|^2}\sum_{(x,y)\in X_i\times X_i}\mu_{R_a}(x,y),$$

$$AIT_a(X_i,X_j)=\frac{1}{|X_i||X_j|}\sum_{(x,y)\in X_i\times X_j}\mu_{R_a}(x,y) \quad (6)$$

그러면 전체적인 상호간의 또는 내부적인 평균 유사성은 각각 다음과 같이 구할 수 있다.

$$OIR_a=\frac{1}{r(d)}\sum_{i=1}^{r(d)}AIR(X_i),$$

$$OIT_a=\frac{1}{r(d)(r(d)-1)}\sum_{i=1}^{r(d)}\sum_{\substack{j=1\\ j\neq i}}^{r(d)}AIT(X_i,X_j) \quad (7)$$

여기서 OIR_a와 OIT_a는 1보다 작으며, 이들을 이용하여 $a\in A$의 중요성을 부여하면

$$DI_a=\frac{OIR_a}{OIT_a}=(r(d)-1)\frac{\sum_{i=1}^{r(d)}AIR(X_i)}{\sum_{i=1}^{r(d)}\sum_{\substack{j=1\\ j\neq i}}AIT(X_i,X_j)} \quad (8)$$

과 같으며 이를 정규화 하여 가중치를 계산하면

$$DI_a=\frac{DI_a}{\sum_{a\in A}DI_a} \quad (9)$$

와 같다.

다. 가중치를 이용한 유사성 척도의 총체화

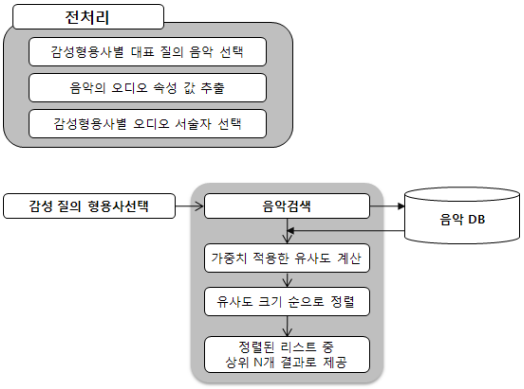
일단 퍼지 유사성을 기반으로 식 (5)와 식 (9) 같이 가중치를 결정하였다면 이들 가중치들은 각각의 특징들 관점에서 유사성 측도를 조합하는데 활용할 수 있다. 만약 가중치가 적용된 합을 통하여 총체적인 유사성을 측정한다면 각각의 유사성을 조합하여

$$R_A(x,y)=\sum_{a\in A}w_aR_a(x,y) \text{ 또는 } \mu_A(x,y)=\sum_{a\in A}w_a\mu_a(x,y) \quad (10)$$

과 같이 계산할 수 있다.

4. 검색 시스템 및 실험 결과

본 연구에서는 (그림 2)와 같은 신송이 등이 제안한 것과 동일한 검색 시스템의 구조를 이용하여, 제안된 가중치 부여방법이 감성기반 음악검색의 성능에 미치는 영향을 실험을 통하여 검토하였다. (그림 2)의 구조는 크게 전처리와 음악검색 두 부분으로 나누어진다.



(그림 2) 검색 시스템 구조

가. 전처리

먼저 감성형용사별 대표 질의 음악은 설문조사를 통하여 선정하였다. 피 실험자 5명이 360곡 중 각 감성형용사를 가장 잘 표현하는 음악 10곡을 선택하여 1~5까지 점수를 부여하였다. 감성형용사별 점수 평균을 내어 가장 점수가 높은 5곡을 각 형용사의 대표곡으로 선정하였다[4]. 음악의 오디오 속성 값은 matlab으로 구현된 프로그램을 이용하여 추출하였다.

마지막으로 검색에 사용할 감성형용사별 오디오 서술자를 선택한다. 서로 대립하는 두 형용사로 이루어진 형용사 쌍에 해당하는 대표곡 10곡(각각의 형용사에 5곡씩)을 이용하여 3장에서 설명한 가중치 결정방법을 적용한다. 퍼지 유사관계를 이용한 러프 근사화 방법에서는 유사도의 문턱치 λ를 0.55로 설정하였다.

<표 1>은 감성형용사 쌍에 대한 MPEG-7 저수준 오디오 서술자들의 중요도를 나타낸다. 표에서 포함정도 열은 신송이 등이 제안한 방법이며, 러프 근사화 방법과 군집 내/군집 간의 유사성 비율은 본 논문에서 제안한 방법이다. 색칠된 부분은 중요도가 높은 서술자로 러프 근사화 방법과 군집 내/군집 간의 유사성 비율 방법은 대체로 일치하는 모습을 보인다.

<표 1> 감성형용사 쌍에 대한 MPEG-7 오디오 서술자의 중요도

	포함정도						러프 근사화						군집 내/ 군집 간 유사성 비율					
	Ann_Cal	Ang_Pea	Ner_Rel	Exc_Sle	Hap_Bor	Ple_Sad	Ann_Cal	Ang_Pea	Ner_Rel	Exc_Sle	Hap_Bor	Ple_Sad	Ann_Cal	Ang_Pea	Ner_Rel	Exc_Sle	Hap_Bor	Ple_Sad
AudioFundamentalFrequency	0.1597	1.0000	0.6588	0.9118	0.4255	0.4255	0.2439	0.1667	0.2381	0.2381	0.2632	0.2041	2.0579	7.6856	3.1147	4.3663	1.7651	2.3365
AudioWaveform	0.6392	0.6373	0.5614	0.7681	0.7132	0.3720	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9918	1.0740	1.0148	1.0695	1.0347	1.1593
InstrumentTimbre	0.4843	0.3608	0.4637	0.2294	0.5740	0.3127	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9421	0.9753	1.0640	1.0937	0.9984	0.9936
AudioSignature	0.8069	0.8627	0.8160	0.8693	0.3078	0.4667	0.0244	0.0167	0.0714	0.0238	0.0263	0.0204	1.2505	1.1837	1.8216	1.3989	1.0805	1.1348
DCOffset	0.4771	0.7627	0.8716	0.4608	0.5799	0.3426	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0458	0.9722	1.1248	1.0640	0.9836	1.0311
AudioHarmonicity	0.1725	0.4771	0.6412	0.7281	0.5869	0.5399	0.0244	0.0167	0.0238	0.0238	0.0263	0.0408	1.1063	1.0892	1.0926	1.1057	1.0830	1.0861
AudioSpectrumBasis	0.8843	0.5686	0.6631	0.6631	0.6631	0.6631	0.2439	0.1667	0.2381	0.2381	0.2632	0.2041	2.2901	2.5765	2.1546	2.2672	2.6262	2.0629
AudioSpectrumCentroid	0.5838	0.7954	0.3814	0.5814	0.6392	0.4951	0.0244	0.1667	0.0476	0.0714	0.0263	0.1429	1.3495	1.7779	1.4833	1.2319	1.3023	1.3517
AudioSpectrumEnvelope	0.6176	0.9588	0.7473	0.9588	0.7569	0.2882	0.0244	0.0500	0.0238	0.0476	0.0526	0.0204	1.1056	1.5144	1.1277	1.4992	1.1664	1.2066
AudioSpectrumFlatness	0.8757	0.9510	0.8748	0.9510	0.3893	0.5644	0.2439	0.1667	0.2381	0.2381	0.2632	0.2041	2.3446	5.7319	4.1206	3.7877	2.0151	2.9963
HarmonicSpectralCentroid	0.4500	1.0000	0.8559	0.8235	0.7892	0.8598	0.0488	0.1667	0.0714	0.0476	0.0526	0.0816	1.2025	2.8454	1.6545	2.7606	1.3985	2.1385
HarmonicSpectralDeviation	0.2965	0.7415	0.7284	0.7889	0.1529	0.3357	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0086	1.1584	1.0895	1.1713	1.0175	0.9937
HarmonicSpectralSpread	0.4810	0.7529	1.0000	0.4834	0.6687	0.3451	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0330	1.1600	1.2233	1.1094	0.9675	1.0126
HarmonicSpectralVariation	0.2034	0.8098	0.5098	0.6255	0.4020	0.3235	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0217	1.1448	1.1097	1.0589	1.0558	1.0452
LogAttackTime	0.3667	0.4431	0.4873	0.1725	0.6191	0.2716	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9421	0.9753	1.0640	1.0937	0.9984	0.9936
SpectralCentroid	0.4422	0.9294	0.4289	0.7794	0.8130	0.8569	0.0244	0.0667	0.0000	0.0238	0.0263	0.0408	1.1204	1.8979	1.1634	1.4672	1.2735	1.6424
TemporalCentroid	0.3910	0.6477	0.8230	0.3893	0.5083	0.1827	0.0000	0.0000	0.0238	0.0000	0.0000	0.0000	0.9868	1.0393	1.2104	1.0191	1.2155	0.9902

나. 음악검색 시스템

제안하는 감성기반 음악검색 시스템은 신송이 등이 사용한 (그림 2)와 같은 구조이다. 사용자가 원하는 감성 형용사를 선택하면 전처리 과정을 통해 선택된 감성에 따른 N 개의 대표 질의 음악을 사용자에게 제시해 준다. 사용자는 대표 질의 음악을 듣고 그 중에 자신이 가장 만족하는 음악 하나를 선택하여 유사한 감성의 음악을 검색하는 감성기반 음악검색을 시작한다. 음악검색은 전처리 단계에서 이미 결정된 서술자와 가중치를 이용하여 식 (10)을 계산한다. 계산된 값을 이용하여 사용자가 선택한 것과 가장 유사한 음악을 찾을 수 있도록 하였다. 유사도 계산 결과는 유사도가 가장 큰 음악부터 내림차순으로 정렬하며 정렬된 음악 중에 상위 10개의 음악을 검색 결과로 사용자에게 제공하였다.

신송이 등이 연구에서 수행한 이전 실험과 동일한 조건을 맞추기 위해 특정 형용사 질의의 대표음악으로 제공되는 음악들은 고정시켜 두었다[7]. 총 10명의 대학생이 실험에 참가하였고 검색에 사용된 음악의 길이는 약 10초이다.

<표 2>는 실험 결과를 표로 나타낸 것이다. 각 감성형용사마다 검색결과로 제시된 10개의 곡 중 피실험자가 만족하는 곡의 개수를 수집하여 평균을 구하였다. 평균검색 결과를 보면 제안된 러프 근사화 방법과 군집 내/군집 간 유사도 비율에 의한 방법이 포함정도를 이용하는 방법보다 상대적으로 좋은 성능을 나타내었다.

<표 2> 가중치 결정 방법에 따른 감성기반 음악검색 결과

가중치 결정방법 / 감성형용사	포함정도	러프 근사화	군집 내/군집 간 유사성 비율
Annoying	2	4.2	3.8
Calm	8	6.0	5.9
Angry	4	5.7	5.3
Peaceful	4	8.7	7.8
Nervous	3	6.4	6.8
Relaxed	3	7.5	7.7
Excited	9	6.7	6.2
Sleep	2	7.0	7.1
Happy	2	6.5	5.5
Bored	3	3.2	3.2
Pleased	4	5.2	4.8
Sad	6	7.0	6.6
평균 검색개수	4.2	6.2	5.9

5. 결론 및 향후 연구 방향

제안된 방법은 내용기반 음악검색을 기반으로한 감성기반 음악검색 구조에서 실험한 결과 평균 검색 개수측면에서 기존 방법보다 좋은 검색 결과를 나타내었다. 가장 좋은 결과를 나타낸 것은 러프 근사화 방법을 이용한 검색이었다. 총 12개의 감성중 Calm 과 Excited를 제외한 10개의 감성에서 기존 방법보다 더 나은 검색 결과를 보였으며 평균적으로도 기존 방법보다 좋은 검색 결과를 나타내었다.

제안된 방법의 감성기반 음악검색 시스템은 6쌍 12개의 감성만 검색 가능한데, 이를 발전시켜 더 다양한 경우에도 사용자의 의도에 맞는 음악을 검색 할 수 있는 시스템을 개발할 예정이다.

참고문헌

[1] Yibin Zhang, Jie Zhou, "A Study On Content-Based Music Classification", IEEE Proc. 7th International Symposium on Signal Processing and Its Applications, vol. 2 , pp. 113-116, 2003.

[2] Erling Wold, Thom Blum, Douglas Keislar, James Wheaton,"Content-Based Classification, Search, and Retrieval of Audio", Multimedia IEEE, Vol. 3, No. 3, pp.27-36, 1996

[3] 박만수, 박철의, 김회린, 강경옥, “MPEG-7 오디오 하위 서술자를 이용한 음악 검색 방법에 관한 연구”, 한국방송공학회 정기총회 및 학술대회, 2003

[4] 신송이, “다중질의 방법을 이용한 감성기반 음악 검색 시스템”, 전북대학교, 석사학위 졸업논문, 2010

[5] Overview of the MPEG-7 Standard (version 6.0), ISO/IEC JTC1/SC29/WG11/N4509.

[6] Z. Pawlak, “Rough sets”, International Journal of Computer and Information Science, Vol. 11, No. 5, 1982

[7] Ekman, P. Emotion in the Human Face ,Cambridge Univ. Press, Cambridge, 1982