스포츠 중계를 위한 자막 인식 시스템 개발

Development of a Video Caption Recognition System for Sport Event Broadcasting

오주현, Juhyun Oh*

요약 메이저리그 야구 중계 등 해외 스포츠 중계제작에서 해결해야 할 문제 중 하나는 MPH(miles per hour)와 같이 영미식 단위로 표시된 자막을 국내 실정에 맞게 km/h 등으로 변환하는 것이다. 이를 위해 중계화면에 표시된 자막영역의 변화로부터 해당 자막이 표시되었음을 감지하고 숫자 정보를 인식하여 이를 국내실정에 맞는 SI 단위로 변환하는 스포츠 자막 인식 시스템을 개발하였다. 변환된 자막은 후단의 문자발생기 (CG) 시스템으로 전달되어 최종적으로 TV 화면에 표시된다. 일반적으로 문자 인식에 주로 사용되는 신경망(neural networks) 기반 방식은 사전에 유사 데이터를 이용한 신경망의 학습(training) 과정이 필수적으로 요구되며, 또한 학습에 사용된 데이터와 다른 모양의 자막이 예고 없이 사용되었을 경우 대처할 수 없다는 단점이 있다. 생방송이라는 사용 환경을 고려하여 새로운 폰트로 제작된 자막에도 신속하게 대처할 수 있는 템플릿 매칭(template matching) 방식을 사용하였다. 여러 가지 실험 영상으로 테스트한 결과 97% 이상의 정확한 인식 결과를 얻었으며, 정확성을 요하는 생방송의 특성상 매칭의 확신도(confidence)가 높지 않은 경우에는 작업자가 판단한 후 핫키를 이용하여 정확한 자막을 출력할 수 있게 하였다.

Abstract A video caption recognition system has been developed for broadcasting sport events such as major league baseball. The purpose of the system is to translate the information expressed in English units such as miles per hour (MPH) to the international system of units (SI) such as km/h. The system detects the ball speed displayed in the video and recognizes the numerals. The ball speed is then converted to km/h and displayed by the following character generator (CG) system. Although neural-network based methods are widely used for character and numeral recognition, we use template matching to avoid the training process required before the broadcasting. With the proposed template matching method, the operator can cope with the situation when the caption's appearance changed without any notification. Templates are configured by the operator with a captured screenshot of the first pitch with ball speed. Templates are updated with following correct recognition results. The accuracy of the recognition module is over 97%, which is still not enough for live broadcasting. When the recognition confidence is low, the system asks the operator for the correct recognition result. The operator chooses the right one using hot keys.

핵심어: *Sport Event Broadcasting, Character Generator, Video Caption, Optical Character Recognition*

*저자 : KBS 방송기술연구소 e-mail: jhoh@kbs.co.kr

1. 서론

국내 선수들의 잇따른 해외 리그 진출로 해외 스포츠의 국내 중계방송 수요가 높아지고 있다. 이와 같은 해외 스포츠 경기 영상을 그대로 국내에 중계할 경우 문제가 되는 것 중의 하나는 한글화되지 않은 자막이다. 스포츠 경기에 있어서 자막은 해당 경기의 진행상황을 한눈에 알 수 있게 하는 중요한 정보이다. 그러나 영미권 스포츠 중계의 자막을 그대로 사용하는 경우 국내 시청자들에게 오히려 혼돈을 줄 수 있으므로, 일반적으로 해외 스포츠 중계 자막 위에 국내에서 비슷하게 제작한 자막을 덧씌워서 방송한다. 야구 중계의 경우, 숙련된 CG 작업자가 현재 점수와 주루 상황 등이 변할 때마다 변경 사항을 수동으로 입력한다.

그러나 이러한 수동 입력으로 해결할 수 없는 문제 중의 하나는 영상에 MPH(miles per hour) 단위로 삽입된 투구속도를 km/h 로 바꾸는 일이다. 투구속도는 중계 자막에 포함된 다른 정보들과 달리 매 투구 시마다 바뀌며, 그때마다 속도에 1.609344 km/mile 을 곱하는 연산이 필요하므로, 투구 시에 눈으로 구속을 확인한 후 암산으로 처리하는 식의 수동 입력 방법은 처리 속도나 정확성 면에서 사용 불가하다. 이와 같은 요구로 영상으로부터 투구속도를 자동 분석하여 km/h 로 환산 표시하는 시스템의 개발 필요성이 제기되었다.

2. 시스템 구성과 템플릿 설정

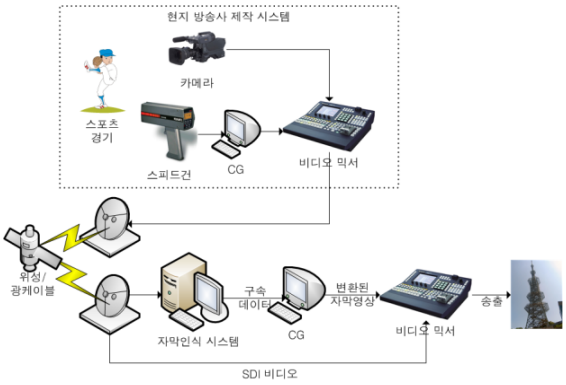

그림 1. 스포츠 중계 제작 시스템

그림 1 은 자막 인식 시스템이 사용되는 중계 제작 시스템의 전체 구성이다. 해외 스포츠 중계 제작이 이루어지는 현지 방송사에서 스피드건 등을 이용해 정보를 측정하고 자막을 삽입하여 송출하면 이를 수신한 국내 방송사의 자막인식 시스템에서 중계화면의 자막을 분석, 숫자를 인식한다. 자막인식 시스템과 RS-232C 로 연결된 후단의 CG 시스템[1]에서 국내 실정에 맞게 변환된 자막을 원 자막 위치에 덧씌워서 최종적으로 송출한다. 자막 인식 시스템에 입력되는 SD 급 SDI (Serial Digital Interface) 비디오를 처리하기 위해 AJA 사의 Xena HS/2Ke 영상 입력 보드를 사용하였다. 자막 인식 소프트웨어는 Microsoft 사의 .NET 기술에 기반하여 개발하였는데, DirectShow.NET 라이브러리[2]를 통해 비디오 입력을 처리하였다. 그림 2 는 자막 인식 운용 소프트웨어의 사용자 인터페이스이다. 그림에서 사용자 인터페이스의 우측 상부는 좌측의 주 화면으로부터 작업자가 지정한 자막영역을 확대하여 표시한다. 투구 속도가 최초로 표시될 때 영상을 캡쳐하여 저장하고, 작업자가 템플릿의 폰트, 크기, 위치 등을 저장된 자막 이미지와 비교하면서 조정한다.

그림 2. 자막 인식 시스템 사용자 인터페이스

이 과정에서 템플릿의 전경/배경 색상은 작업자가 신속하게 선택하는 것이 쉽지 않으므로, RGB 3 차원 컬러 공간에서 K-means 클러스터링 알고리즘[3]을 사용하여 자동으로 선택되도록 하였다. 템플릿 조정 작업이 끝나서 '적용' 버튼을 누르면 'MPH' 자막 영역이 비트맵으로 메모리에 저장되고, 작업자가 설정한 폰트, 크기, 위치 등의 정보를 이용하여 그림 3 과 같이 0 에서 9 까지의 템플릿이 생성되어 자막인식을 실행할 준비가 완료된다.

그림 3. 생성된 템플릿 이미지

3. 자막 인식

이미지와 비디오에서 자막의 존재 여부, 자막 영역의 위치 결정, 자막 추적, 자막 추출(segmentation), 인식

결과를 이용한 자막 이미지의 개선 등은 [4]에 잘 정리되어 있다. 본 논문에서 자막의 존재는 2 장에서 언급한 바와 같이 사용자가 첫 번째 투구 샷을 저장한 다음 자막 영역을 수동으로 설정하도록 하였다. 또한 자막의 존재 여부는 'MPH' 자막 위치를 매 프레임마다 기 저장된 'MPH' 템플릿과 비교하여 검출하도록 하였다. 자막을 인식한 후에는 마찬가지로 'MPH' 자막 위치를 템플릿과 비교하여 자막이 아웃되었는지 여부를 판단하였다.

일반적으로 문자와 숫자 인식에는 다층 퍼셉트론(multi-layer perceptron) 등 신경망(neural networks) 기반 방식이 주로 사용된다[5,6]. 그러나 이 방법은 인식 시스템 활용 이전에 유사 데이터를 이용한 신경망의 학습(training) 과정이 필수적으로 요구된다. 또한 학습에 사용된 데이터와 전혀 다른 모양(폰트, 크기, 색상 등)의 자막이 사용되었을 경우 대처할 수 없다는 단점이 있다.

본 시스템이 사용될 중계제작 현장이라는 사용 환경을 고려할 때, 학습을 위한 시간과 공간 등의 제작 리소스가 절대적으로 부족하고 메이저리그 현지 방송 제작사로부터 매 경기마다 어떤 모양의 자막이 사용될 것인지에 관한 정보를 사전에 수신하는 것이 불가능하므로, 사전 학습이 필요한 신경망 기반 방식은 배제하였다. 대신 사전 처리 과정이 필요 없고 새로운 글꼴로 제작된 자막에도 신속하게 대처할 수 있는 템플릿 매칭(template matching) 방식이 사용되었다.

그림 4 는 일반적인 메이저리그 중계 화면이며, 그림 5 는 다양한 스타일의 자막과 애니메이션 효과를 보여 준다. 자막 숫자를 인식하기 전에 우선 화면에 투구속도 자막이 존재하는지 판단할 필요가 있는데, 이는 최초 템플릿 생성 과정에 저장했던 'MPH' 이미지를 화면 상에서 검출함으로써 이루어진다. 그러나 자막이 등장할 때 디졸브(dissolve)나 와이프(wipe) 등의 애니메이션 효과가 들어갈 뿐 아니라, 영상 신호가 비월주사(interlaced scan)를 사용함으로써 투구속도 자막이 완전히 나타나기도 전에 'MPH' 이미지를 검출하는 경우가 발생할 수 있다.

그림 4. 일반적인 메이저리그 중계 화면

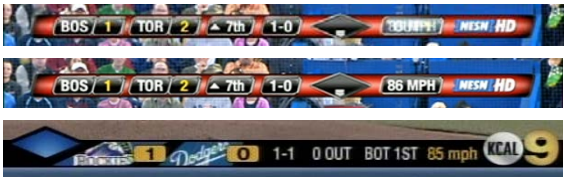

그림 5. 다양한 자막 스타일과 애니메이션 효과

즉 그림 5 의 첫 번째나 세 번째 경우처럼 자막이 완전하게 등장하지 않은 상황에서 자막 인식을 시도하게 됨으로써 인식률을 크게 저하시킬 수 있다. 이와 같은 문제를 막기 위해 최소 두 프레임 이상 'MPH' 이미지가 안정적으로 검출될 경우에만 자막 인식을 시도하도록 하였다.

투구속도 자막영역이 성공적으로 검출되면 각각의 템플릿에 대하여 (1)의 MSD(Mean Squared Difference)를 계산하여 MSD_0 에서 MSD_9 중 가장 낮은 MSD 를 가진 템플릿의 숫자를 선택한다.

$$MSD_i = \sum_{u,v}[T_i(u,v) - I(u,v)]^2 \qquad i = 0,1,2,\ldots,9 \quad (1)$$

여기에서 T_i는 최초에 작업자가 설정한 템플릿 이미지, I 는 질의된 투구속도 자막 이미지, u, v 는 이미지 좌표이다. 이와 같은 템플릿 매칭은 앞에서 언급한 'MPH' 이미지 검출에도 동일하게 사용되며, 'MPH' 이미지 검출에는 미리 정의한 문턱값(threshold)이 사용된다. 그림 6 에 전체 자막 인식 프로세스의 순서도를 나타내었다.

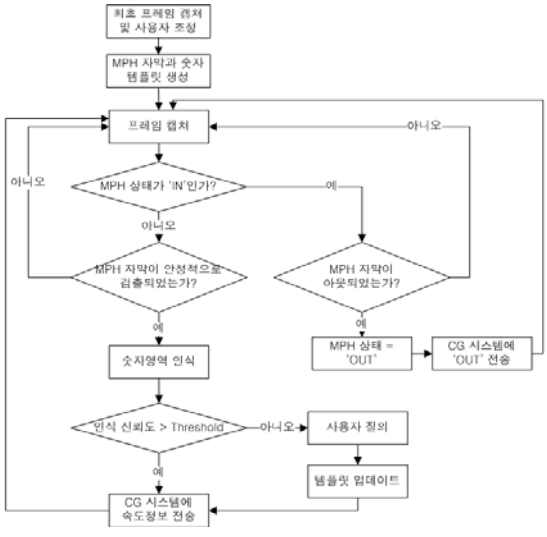

그림 6. 자막인식 프로세스 순서도

표 1. 실험영상의 자막 인식 정확도

경기	실험영상의 투구 수	인식 실패	정확도
LA-CO	72	2	97.2%

MIL-SD	60	0	100%
BOS-TOR	34	1	97.1%
SEA-NY	14	0	100%

방송에 적용하기 전 테스트 영상으로 실험한 결과는 표 1 과 같다. 대체로 100%에 가까운 인식률을 보이고 있으나, 자막의 크기가 작고 두 숫자 간 거리가 충분하지 않은 경우 '5', '6', '8' 등 비슷한 모양의 숫자에서 오인식이 발생하였다. 정확성이 요구되는 생방송의 특성상 작업자가 추가 확인을 할 수 있도록 하였다. 그림 7 과 같이 자막인식 모듈에서 템플릿 매칭 결과 1 위와 2 위의 차이가 충분히 크지 않은 경우, 후단의 CG 시스템으로 불확실한 구속 정보를 바로 넘기는 것이 아니라, 작업자에게 질의하도록 한 것이다. 작업자는 대화 상자가 뜨면 바로 키보드의 '1' 또는 '2' 키를 누름으로써 두 가지 보기 중 하나를 선택할 수 있다.

그림 7. 사용자 질의 화면

이와 같이 작업자에 의해 선택된 숫자 이미지는 그 정확성을 신뢰할 수 있으므로 이를 이용해 템플릿을 업데이트한다. 해당 숫자가 다시 등장했을 때 최초 작업자가 생성한 템플릿이 아니라 저장된 실제 자막 이미지 템플릿과 비교하므로 정확성을 향상시킬 수 있다. 그러나 이와 같이 사용자 생성 템플릿과 실제 자막에서 추출된 템플릿이 혼재할 경우, 상당수 실제 숫자와 관계없이 자막 추출 템플릿이 우선적으로 선택되었다. 이는 사용자 생성 템플릿의 경우 K-means 방법으로 선택한 폰트 컬러가 실제 자막 이미지와 일치하지 않아서 상대적으로 컬러와 히스토그램 분포가 실제 자막 이미지와 일치하는 자막 추출 템플릿이 선택되는 것으로 판단된다. 따라서 자막 이미지에서 추출한 템플릿이 존재하는 경우 다음과 같은 조건을 추가적으로 만족할 경우에만 자막 추출 템플릿을 선택하도록 하였다.

추가 조건 1: 자막 이미지와 자막 추출 템플릿의 MSD 〈 Threshold1

추가 조건 2: 자막 추출 템플릿만 비교 결과 1 위와 2 위의 차이 〈 Threshold2

표 2 는 자막인식 모듈에서 사용된 여러 가지 상수 값과 그에 대한 설명이다. 이 상수들은 별도의 XML 파일에 저장되어 있으며 인식 성능을 개선하기 위해 수정할 수 있다.

표 2. 자막 인식에 사용된 상수 값들

이름	값	의미
ThrMPH	0.05	'MPH' 영상을 검출하기 위한 문턱값 조건: MSD 〈 ThrMPH
ThrStableMPH	0.05	Fade-in, Interlacing 등 투구속도 등장 시 발생할 수 있는 왜곡에 대비해 안정적인 자영상 검출을 위한 문턱값
ThrGap	0.01	템플릿 매칭에서 1위와 2위의 차이를 비교. 차이가 ThrGap보다 작으면 사용자에게 질의
ThrStored	0.05	사용자가 저장한 템플릿과의 비교 (MSD) 결과가 일정 값 이하일 때 수용
ThrStoredGap	0.0005	사용자 저장 템플릿 1,2위의 차이가 일정수준 이상일 때 수용

4. 결론

증가하는 해외 스포츠 중계 방송 요구에 대응하여 스포츠 자막인식 시스템을 개발하였다. 템플릿 매칭에 기반한 방식으로 생방송에 적용하기 어려운 사전 학습 과정 없이도 우수한 인식률을 얻을 수 있었다. 문자 인식과 함께 인식의 신뢰도를 감안하여 필요한 경우 사용자 입력을 받아 오류 없는 중계방송이 가능하도록 하였다. 본 시스템은 KBS 의 메이저리그 야구 중계방송에서 효과적으로 사용되었으며, 자막 정보를 이용한 경기 요약과 메타데이터 생성 등 다양한 용도로 활용하기 위한 기반 기술이 될 수 있을 것으로 기대된다 [7]. 향후 사용자 개입을 최소화하기 위해 인식률을 높이는 추가 연구를 수행할 계획이다.

참고문헌

[1] 박근수 외, "디지털 제작 시스템 연구", KBS 방송기술연구소 2006 년 연구보고서, pp. 219~267, 2006.

[2] Thomas Scheidegger, "DirectShow.NET", http://www.codeproject.com/KB/directx/directshownet.aspx, 2002.

[3] Sanjiv K. Bhatia, "Adaptive K-Means clustering", Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, 2004.

[4] Keechul Jung, Kwang In Kim, and Anil K. Jain, "Text information extraction in images and video: a survey", Pattern Recognition, Vol 37, Issue 5, May 2004, pp. 977~997.

[5] D. Chen, "Text Detection and Recognition in Images and Video Sequences", Ph.D. thesis, IDIAP, Switzerland, 2003.

[6] M. Egmont-Petersen, D. de ridder, and H. Handels, "Image processing with neural networks a review", Pattern Recognition, 2002, pp. 2279-2301.

[7] 유기원, 허영식, "자막 정보를 이용한 야구경기 비디오의 자동요약 시스템", 방송공학회 논문지, 2002, 제 7 권 제 2 호 pp. 107~113.